Electroencephalograph (EEG) Signal Processing Techniques for Motor Imagery Brain Computer Interface Systems

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

by

THIYAM DEEPA BEETA

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

June, 2018

DECLARATION

I here by declare that the thesis entitled "EEG Signal Processing Techniques for Motor Imagery Brain Computer Interface Systems" submitted by me, for the award of the degree of *Doctor of Philosophy* to VIT University, is a record of bonafide work carried out by me under the supervision of Dr. E. R. Rajkumar, Associate Professor, SENSE, VIT University, Vellore (currently working as Senior Architect, Robert Bosch Engineering and Business Solutions Private Limited, Bangalore) and Dr. Sergio A. Cruces Alvárez, Associate Professor, Signal Processing Group, University of Seville, Spain.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in these institutes or any other institute or university.

Place: Vellore

Date: 11/06/2018

Signature of the Candidate

CERTIFICATE

This is to certify that the thesis entitled "EEG Signal Processing Techniques for Motor Imagery Brain Computer Interface Systems" submitted by Ms. THIYAM DEEPA BEETA, School of Electronics Engineering (SENSE), VIT University, Vellore for the award of the degree of *Doctor of Philosophy* and Ingeniería Automática, Electrónica Y de Telecomunicacíon, University of Seville, Spain for the award of the degree of *Doctor*, is a record of bonafide work carried out by her under our supervision, as per the VIT and University of Seville code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in these institutes or any other institute or university. The thesis fulfills the requirements and regulations of the Universities and in our opinion meets the necessary standards for submission.

Place: Vellore

Date: 11/06/2018

Signature of the VIT Guide

(Dr. E. R. Rajkumar)

Signature of the University of Seville Guide

(Dr. Sergio A. Cruces Alvárez)

UNIVERSIDAD DE SEVILLA
DEPARTAMENTO DE TEORÍA DE LA SEÑAL Y COMUNICACIONES

ABSTRACT

Brain-Computer Interface (BCI) system provides a channel for the brain to control external devices using electrical activities of the brain without using the peripheral nervous system. These BCI systems are being used in various medical applications, for example controlling a wheelchair and neuroprosthesis devices for the disabled, thereby assisting them in activities of daily living. People suffering from Amyotrophic Lateral Sclerosis (ALS), Multiple Sclerosis and completely locked in are unable to perform any body movements because of the damage of the peripheral nervous system, but their cognitive function is still intact. BCIs operate external devices by acquiring brain signals and converting them to control commands to operate external devices. Motor-imagery (MI) based BCI systems, in particular, are based on the sensory motor rhythms which are generated by the imagination of body limbs. These signals can be decoded as control commands in BCI application. Electroencephalogram (EEG) is commonly used for BCI applications because it is non-invasive. The main challenges of decoding the EEG signal are because it is non-stationary and has low spatial resolution. The common spatial pattern algorithm is considered to be the most effective technique for discrimination of spatial filter but is easily affected by the presence of outliers. Therefore, a robust algorithm is required for extraction of discriminative features from the motor imagery EEG signals.

This thesis mainly aims in developing robust spatial filtering criteria which are effective for classification of MI movements. We have proposed two approaches for the robust classification of MI movements. The first approach is for the classification of multiclass MI movements based on the thinICA (Independent Component Analysis) and mCSP (multiclass Common Spatial Pattern Filter) method. The observed results indicate that these approaches can be a step towards the development of a robust feature extraction for MI based BCI system.

The main contribution of the thesis is the second criterion, which is based on Alpha-Beta logarithmic-determinant divergence for classification of two class MI movements. A detailed study has been done by obtaining a link between the AB log det divergence and CSP criterion. We propose a scaling parameter κ to enable similar way for se-

lecting the respective filters like the CSP algorithm. Additionally, the optimization of the gradient of AB log-det divergence for this application was also performed. The Sub-ABLD (Subspace Alpha-Beta Log-Det divergence) algorithm is proposed for the discrimination of two class MI movements. The robustness of this algorithm is tested with both the simulated and real data from BCI competition dataset. Finally, the resulting performances of the proposed algorithms have been favourably compared with other existing algorithms.

ACKNOWLEDGEMENT

I am very grateful to my supervisor **Dr. E. R. Rajkumar**, Associate Professor, SENSE, VIT University (currently working as Senior Architect, Robert Bosch Engineering and Business Solutions Private Limited, Bangalore) for being a constant support throughout my PhD and motivating me to take up challenges. I am also thankful to him for giving me many wonderful opportunities. I owe my sincere gratitude to my co-supervisor **Dr. Sergio A. Cruces**, Associate Professor, Signal Processing Group, University of Seville for introducing me to this interesting field of signal processing. I am very grateful to him for his efforts to teach me right from scratch till the very end and also for being there to answer all my questions. This work would not have been possible without their support.

I would like to thank the **Chancellor** of VIT University, Vellore, the **Rector** of University of Seville, Spain and the management of both the universities for allowing me to carry out this co-directed PhD and also for providing me infrastructural facilities and resources needed for my research work.

I am also thankful to the **Vice chancellor**, **Dean Academic Research**, VIT University, Vellore for all the administrative help and support. I take this opportunity to express my thanks to **VIT University**, **Erasmus Mundus European Commission** and the **Spanish Government** for providing funds for my research work.

I would also like to acknowlegde my doctoral committee members, **Dr. Venkatesh B.**, Professor, Department of Engineering Design, IIT madras, Chennai, **Dr. Sriraam N.**, Head of Deparment, Department of Medical Electronics, M. S. Ramaiah Institute of Technology, Bangalore and **Dr. Arulmozhivarman P.**, Dean, School of Electrical Engineering, VIT University, Vellore for their kind and timely suggestions to improve the quality of my research work.

I am indebted to my friends **Rucha**, **Poulami**, **Lalitha** and **Cassandra** for always lending a helping hand in times of my need. I would also like to extend my thanks to my other friends, colleagues and room-mates who have always been there for me, especially during my hard times.

My deepest gratitude to my colleagues **Juan Antonio**, **Irene Santos**, **Sunny Dayal**, **Maria Jose**, **Irene Fondon**, **Pablo Anguilera**, **Uxi** at the University of Seville,

for making my stay wonderful and for always being there for me. My stay in Seville would have been difficult without their support. A special thanks goes to **Javier** for helping me with the experimental studies.

My heartfelt gratitude goes to my **family** and to **Amarjit**, for their constant encouragement and moral support along with patience and understanding throughout my work. Lastly, I thank **God** and my heavenly **Father** for supporting and blessing me to complete this degree.

Place: Vellore

Date: 11/06/2018

THIYAM DEEPA BEETA

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF TERMS AND ABBREVIATIONS

LIST OF NOTATIONS

CHAPTER 1

Introduction

1.1 Overview

The Brain-Computer Interfacing (BCI) system aims at building the bridge between the brain and the computer. The brain produces electrical and magnetic signals while performing different functions. These signals can be recorded using different techniques. Electroencephalograph (EEG) is the most commonly used method for measuring the electrical activity of the brain in BCI applications because of its non-invasive characteristics. BCI system enables a disabled person to operate other assistive devices like a wheelchair or robotic arm by using brain signals. The Motor Imagery (MI) based BCI systems are considered as the most preferable BCI systems. They use MI EEG signals as control commands for external devices without using the peripheral nervous system. Such systems can potentially serve as a communication aid for the people suffering from Amyotrophic Lateral Sclerosis (ALS), Multiple Sclerosis and completely locked-in. The main difference and advantage of BCIs over other assistive devices are the non requirement of any form of muscle control. Unfortunately, the performance accuracy of current BCIs is still very low restricting to use them out of the laboratories. The main reason for the hinder of performance is due to the non-stationary nature of the EEG signals. As a result of this, the signal properties not only changes from person to person but also from trial to trial which gives more challenges in data analysis. In addition to this difficulty, the presence of artifacts such as eye movements, muscle activities and improper placement of electrodes added more challenges in the EEG signal processing. Furthermore, the performance of the BCI system also decreases with the increased in the number of motor imagery movements. Although, the artifacts and non-stationarity cannot be fully removed, a robust signal processing algorithm can be used for better signal analysis and high classification accuracies. Several approaches have been proposed for designing a robust signal processing unit but still, the gap is large to highlight the BCI system for real time applications. Therefore, this motivates the necessity of developing a more robust algorithm for MI movements classification.

1.1.1 Objective of the Thesis

This thesis focuses on the study of different spatial filtering methods used for classification of MI movements and developing robust algorithms based on it for the application of EEG based MI-BCI system. The first objective is to understand and analyze the various filtering methods in this area. Another objective can be divided into two sections. The first section is to propose a new criterion for the classification of multiclass motor imagery movements. The second section, which is the main contribution of this thesis, will formulate a new criterion based on the Alpha-Beta Logarithmic-Determinant divergence (AB Log-Det) for discrimination of two class MI movements. The list of contributions is the following:

- A novel criterion for classification of multiple MI movements is proposed, it is based on the extension of the Thin Independent Component Analysis (ThinICA) method for blind source separation.
- ThinICA-Common Spatial Pattern (CSP) algorithm is proposed combining the CSP and the extension of the ThinICA method for multiple MI movements classification.
- The relation between CSP and AB Log-Det divergence is determined.
- The scaling factor κ, which provides the equal solution between the CSP and AB Log-Det divergence is obtained.
- A novel regularized criterion is proposed based on AB Log-Det divergence.
- The optimization of AB Log-Det divergence with proper gradient function is presented
- The Subspace-Alpha Beta Log-Det Algorithm (Sub-ABLD) algorithm is proposed to address the robust features extraction problem in MI-BCI systems.

1.1.2 Structure of the Thesis

This thesis consists of nine chapters and four appendices. This chapter presents the introduction, motivation, problem statement, objectives and structure of the thesis.

In Chapter 2, the basic anatomical and physiological details of the human brain, different types of brain rhythms and the various types of EEG artifacts are introduced. The BCI and MI-based BCI as well as the commonly used filtering and classification algorithms for the discrimination of motor imagery EEG signals are also discussed.

In Chapter 3, the existing effective spatial filtering approaches are reviewed in details. The popular CSP algorithm, variants of CSP and other different spatial filtering approaches for discrimination of motor imagery movements are described.

In Chapter 4, the simplification of CSP objective function as Rayleigh quotient is shown. Moreover, different CSP based approaches such as discriminative CSP and ICA corrections to CSP based on the existing techniques are presented.

In Chapter 5, the extension of existing Thin-ICA criterion is presented. The maximization of the proposed criterion is done for the classification of multiple class MI movements. The Thin-ICA CSP algorithm combines the proposed criterion with the solution of the multiclass CSP algorithm. The performance results using BCI competition dataset are also presented.

In Chapter 6, the field of AB Log-Det divergence is introduced for BCI applications. The optimization of this divergence is performed and its interpretation of CSP is obtained. The scaling parameter that provides the equivalent solution between the AB Log-Det divergence and CSP is presented. The gradient of AB Log-Det divergence is computed and validated. The robustness of this divergence based on α and β is shown.

In Chapter 7, the criterion based on AB Log-Det divergence is proposed for addressing the problem for discrimination of two class MI movements. Sub-ABLD algorithm is proposed by optimizing this criterion. This proposed criterion considers both the within class and between class divergence. The algorithm is evaluated using both artificial and BCI competition datasets.

In Chapter 8, the simulations of the both the proposed algorithms i.e. ThinICA-CSP and Sub-ABLD algorithm are compared with the other baseline algorithms. The study of the performance of Sub-ABLD algorithm in different scenarios are also presented and the performance results are discussed.

In Chapter 9 gives the final conclusion of this thesis and some future research ideas are presented.

CHAPTER 2

Brain Computer Interface Background

The human brain, the most complicated organ of the human body, has the ability to control the other parts of the body. It produces electrical and magnetic signals while performing different functions. These signals can be recorded using different techniques. EEG was first discovered by Hans Berger (Berger, 1929) and since then it has become very popular for the analysis and diagnosis of various brain disorders. It involves measuring the electrical activity of the brain by using the scalp electrodes. The EEG system has certain advantages over the other measurement techniques such as non-invasive, high temporal resolution, low cost and portable features. Besides being used in the clinical applications, EEG was later used for a man-computer interface application. This was first introduced by Jacques J. Vidal in 1973 (Vidal, 1973). Initially, BCI was based on the neurofeedback which requires a long training process (Spilker et al., 1969). However, with the advances in the signal processing techniques, the current BCI has the ability to decode the EEG signals as a control commands for the computer. Moreover, it can also adapt based on the user's intention.

Recently, BCI (Dornhege, 2007) has gained lots of interest in neuroscience and rehabilitation engineering. It provides an alternative pathway to control the external devices with the brain signals without using the peripheral nervous systems. Hence, this feature makes BCI one of the most favourable choices in the field of neuro-rehabilitation. A person suffering from ALS or completely locked in cannot perform the movement of the body limbs and muscles efficiently due to the weakness in the muscles. Here comes the role of BCI by providing alternative communication channels using the brain signals. Besides medical applications, BCI has also been used for the development of games (Krauledat et al., 2009; Bonnet et al., 2013) and many other non-medical applications (Van Erp et al., 2012).

2.1 Structure of The Brain

The human brain, owing to its extent of physiological control of the human body, has been the subject of analysis, modelling and recently, rehabilitation. Anatomically, the

brain is divided into the cerebral cortex, the cerebrum and the brain stem. Alternatively, the brain structures can also be classified into the forebrain, midbrain and hindbrain regions. Each part of the human brain is responsible for performing a different function. The frontal lobe is responsible for thinking, concentration, language and personality. The

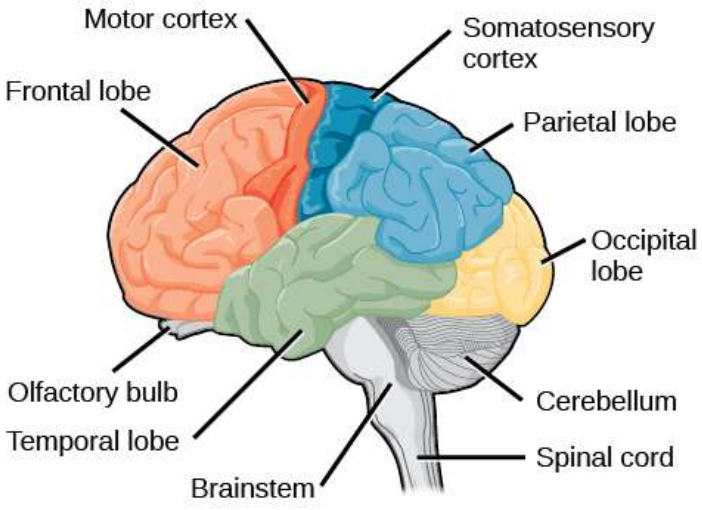


Fig. 2.1 Structure of the brain

Source:https://opentextbc.ca

temporal lobe is the site for the auditory reception, memory and information retrieval. The occipital lobe is functional for visual reception and interpretation. The parietal lobe processes sensory inputs, orientation of the body and is also responsible for sensory discrimination. Voluntary motor tasks are controlled by the motor cortex. The cerebellum coordinates the voluntary movements and also controls them whereas the brainstem is responsible for activities like breathing, digestion and control of the heart. The locations of the cerebral lobes and the other structures of the brain are depicted in Fig. 2.1.

2.2 Motor Control

Within the motor cortex itself, several areas are responsible for the various aspects of motor activity as depicted in Fig. 2.2. The Premotor Cortex Area (PMA) is responsible for the sensory guidance required for movement. The Supplementary Motor Area (SMA) takes into consideration all the preparatory aspects of movement and initiates movement whereas the Primary Motor Cortex (PMC), which is also represented as M1, is the area that is actually responsible for the execution of this motor activity. Within M1 area discrete somatotopic organization can be observed, i.e., different regions in M1 are responsible for movement in specific regions of the body. The neurons are aggregated in

these specific areas for each organ. Such organization but in a slightly broader sense has also been found to exist in the SMA region of the motor cortex.

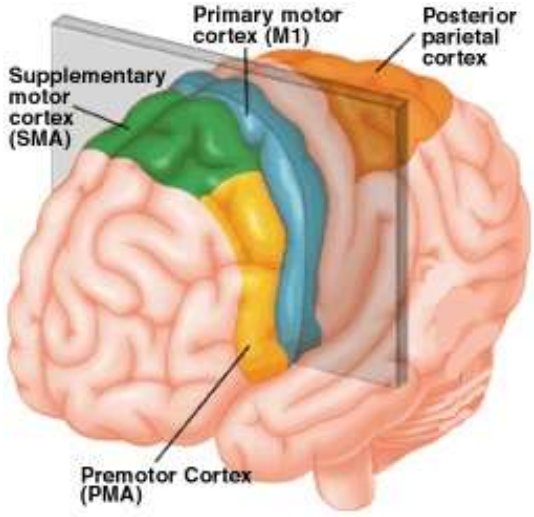


Fig. 2.2 Main cortical regions involved in the motor system

Source:http://brainconnection.brainhq.com/wp-conten

2.3 EEG Signal Acquisition

EEG (Berger, 1929) is a method of recording the electrical signals produced in the brain. EEG signals can be of different types based on the cause. Spontaneous EEG signals, as the name suggests, do not need an external trigger. Evoked potentials, on the other hand, are EEG signals which are obtained when an activity or cognitive task is being performed deliberately. Event potential is EEG observed during the occurrence of a particular event. EEG signals are recorded by placing the electrodes on the scalp. Scalp EEG and intracranial EEG directly records the EEG signals from the exposed area of the brain. In scalp EEG, which is the more common of the two methods, the signal acquisition procedure is more convenient for the patient as compared to intracranial EEG. On the other hand, the latter yields more accurate and cleaner signals than the former. The placement of electrodes on the scalp for scalp EEG is determined by the 10-20 system of electrodes accepted internationally as shown in Fig. 2.3 (Jasper, 1958). According to this system, the anatomical location of every electrode is defined in terms of percentages (10 or 20%) of the distance between two landmarks: Nasion and Inion. Nasion is located at the beginning of the nose between the eyes, while Inion lies at the base of the skull. The electric potentials of these electrodes are obtained by reference to

a fixed reference electrode; this difference between the two potentials is used for further processing. The representation of electrodes in the 10-20 system is done using numbers and alphabets.

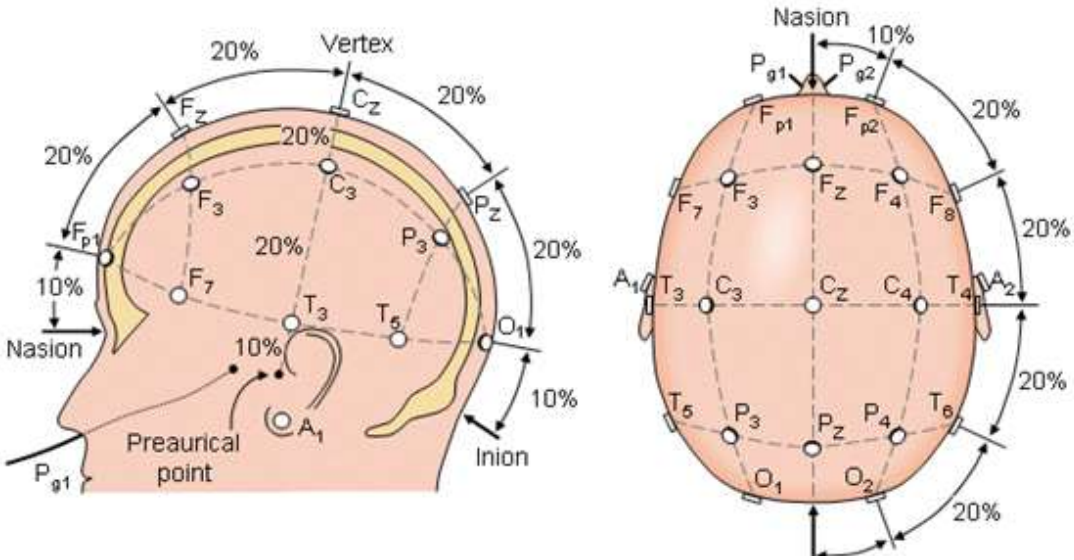


Fig. 2.3 Placement of EEG electrodes (a): Lateral view (b): Top view

Source: http://www.bem.fi

The even numbers and odd numbers indicate the right hemisphere and the left hemisphere respectively, while the alphabets (F, C, O, P, T, Fp) denote the locations (Frontal, Central, Occipital, Parietal, Temporal, Frontopolar respectively). The visualization montage may vary; some types are bipolar, referential and Laplacian montage. Bipolar montage is useful for measuring adjacent differences in potential and thus, better suited to observe localized differences. Referential montage is obtained with respect to a single electrode. Laplacian montage is visualized as the difference between an electrode and the average of its neighbours. The purpose of the EEG defines the necessary montage.

2.3.1 EEG Electrodes

The first electrical activity of the human brain was recorded using the scalp electrodes and a galvanometer by Hans Berger in 1924. Thereafter, many changes have been taken place in the acquisition of brain signals using electrodes. The major challenges in the acquisition of EEG signals are the localization of the electrode montage and to maintain an acceptable level of skin impedance. Initially, in order to decrease the skin impedance, a part of the outer skin is removed. Soon this was replaced with the used of Ag/AgCl electrodes which was minimally invasive. Despite being less invasive the Ag/AgCl

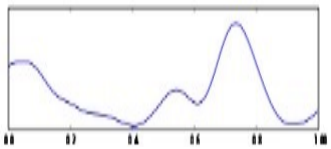
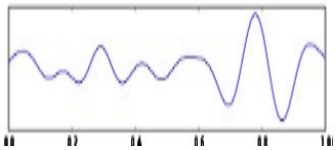
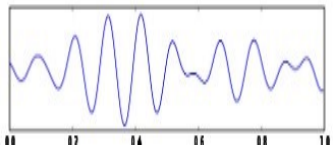
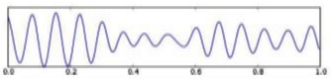
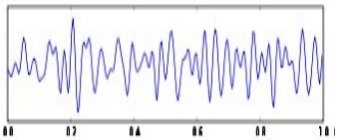
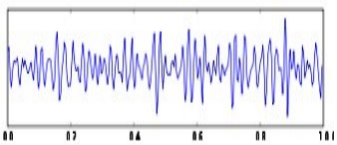
electrode utilizes a sticky electrolyte gel to provide better electrode-skin impedance by making the scalp and hair dirty. Moreover, the impedance of the wet electrodes deteriorates as the acquisition time increases; hence it is not suitable to use for long term acquisition (Gargiulo et al., 2010). Recently, to overcome the disadvantages of using wet electrodes, several approaches have been proposed for the designing on dry electrodes based on various approaches. The dry electrodes can be designed in spiky form where the electrode surface consists of an array of spikes and can be placed directly in contact with the scalp. The spike can be of different scales such as nanometres, micrometres and millimetres. The author of (Griss et al., 2002) proposed a microneedle electrodes that are suitable for real time and long term monitoring. Later a multiwalled carbon nanotube arrays were presented in (Ruffini et al., 2008). A microtip 4×4 dry electrodes was also presented which can record as well as can perform electro-tactile stimulation. Other studies include the designing of dry electrodes using Microelectromechanical systems (MEMS) (Chiou et al., 2006) but the length of the electrodes is not enough for acquiring the signal on the hairy area. The nano and microneedle provide low electrode impedance, less artifacts due to movement and comfortable for long term measurement. Inspite of this, it is not cheap to produce, invasive and fragile at times. To overcome this problem, the author of (Salvo et al., 2012) proposed a non-invasive 3D printer in the micrometric scale which can be reused many times. Another author developed a low cost polymer silver coated electrodes and evaluated with for the BCI applications (Grozea et al., 2011). However, it was invasive and unsuitable for long term used, therefore a non-invasive electrode was proposed in (Liao et al., 2011). As discuss earlier that the nano and microelectrodes may causes loss of contact due to the presence of the hair. Other group of researches proposed a capacitive electrodes where the probe is placed far from the scalp (Harland et al., 2002), (Sullivan et al., 2007), (Oehler et al., 2008).

2.3.2 EEG Rhythm

Although EEG signals appear to be random signals, different rhythms can be observed within their frequencies corresponding to specific mental states from deep sleep to wakefulness (Blume, 1999). The waves and their corresponding frequencies which occur in the typical human EEG signals are represented in Table 2.1 below.

Alpha and Beta waves are present in wakefulness, where the former can be observed in a more relaxed state during wakefulness as compared to the other. Theta waves are observed during sleeping while delta waves are characteristic of deep sleep. Another rhythm, rarely observed, is the *Gamma* wave obtained in high-frequency regions of above 30 Hz and present in situations of high energy and focus. The *mu* rhythm overlaps with the *alpha* rhythm but is generated only when imagining body movements. The mu rhythm is very important for detecting body movements in the BCI applications. The

Table 2.1 Brain rhythms

Name	Band (Hz)	Characteristics	Location	Waveform
Delta (δ)	$0.1-4$	Very low frequency waves. For adults observed at the time of deep sleep. Common in infants and children observed during wakefulness	frontally in adults, posteriorly in children	
Theta (θ)	$4-8$	Present in adults during drowsiness or an idling state and also normally observed in young children	various locations that are not involved in any apparent task	
Alpha (α)	$8-13$	Caused by closing eyes, relaxation and attenuate as one becomes involved in some mental task	posterior regions, occipital and temporal cortex	
Mu (μ)	8 - 13	Effected by actual movement, motor imagery or stimulation	sensorimo-tor cortex	
Beta (β)	$13-30$	Correlated with active thinking, focus, stress, and an alert state	frontal regions, somatosen-sory cortex	
Gamma (γ)	>30	Present during the highly attentive states of consciousness and perception which involves higher mental activity	various locations	

Beta rhythm is generated when planning to execute a movement. Therefore, these two rhythms are most important for MI-based BCI system.

2.3.3 EEG Artifacts

Signals which are not generated by the brain but are visible in the EEG recording are termed as EEG artifacts. These artifacts interfere with the analysis of the EEG and hence, the knowledge of their source and nature is essential for their removal (Urigüen and Garcia-Zapirain, 2015). Typically, the EEG signals are 10 to 100 mV in amplitude. The main types of EEG artifacts are physiological and non-physiological. Physiological artifacts originate from the other regions of the subject's own body and are often closer in magnitude to the EEG signals than non-physiological artifacts.

2.3.3.1 Physiological Artifacts:

The different types of physiological artifacts present in the EEG signals are listed below:

Electromyograph (EMG): EMG is the most common type of artifacts that is picked up during EEG acquisition, due to movement or muscle twitching. However, since its duration is smaller and morphology distinct, it is easily separable except in cases such as Parkinsons or Huntingtons diseases where it is not so easily identifiable due to the high frequency of occurrence.

Electrooculograph (EOG) (Urigüen and Garcia-Zapirain, 2015): Eye movements are generally visible as a part of EEG. They occur due to blinking or the axial movement of the eyeballs. Yet, these movements can be beneficial when dealing with sleep EEG where eye movements can be corresponding to the nature of the EEG waves.

Tongue movements: Similar with the eyes, the tongue also can interfere with the EEG signals due to chewing, biting and other tongue movements, especially those involving the tip of the tongue. These artifacts can be removed by recording these movements separately and then eliminating them from the recorded signals.

Electrocardiograph (ECG): ECG artifacts are observed in EEG signals when there are electric changes in the heart, which occurs either because the subject is obese, has heart disease or when the inter-electrode distances are high.

Pulse: Pulse interferes with EEG signals when the electrodes are placed directly on the surface of a pulsating vessel on the skull.

2.3.3.2 Non-Physiological Artifacts:

In addition to the physiological artifacts, non-physiological artifacts also affect the EEG signals. The common non-physiological artifacts present in EEG signals are described below.

Power-line interference: The interference caused by the frequency of the electric supply can be eliminated as the frequency is specific and known beforehand. This changes the baseline of the EEG signal, hence it needs to be removed.

Artifacts due to electrodes: These may occur due to the sudden movement or popping of an electrode. This is easily identified on the EEG signal as a sudden vertical transient. The improper contact of the electrodes with scalp can also lead to the distortion of the EEG signals

Environmental conditions: Movements of other people in the surroundings may be the cause of artifacts in the EEG signals. Electrical interference from other devices may produce artifacts. Respirators and other equipment can also alter the EEG recordings.

Some of the artifacts can be easily removed by filtering the EEG signals with appropriate filters whereas other artifacts are removed by using other signal decomposition techniques such as Independent Component Analysis (ICA), Principal Component Analysis (PCA) etc.

2.3.4 Event Related Synchronization and Desynchronization (ERS/ERD)

During movement, two kinds of signals are generated. The low-frequency signal generated in preparation of movement is termed as Movement Related Potential (MRP). The high frequency signals are termed as Event-Related Synchronization/ De-synchronization (ERS/ERD) (Pfurtscheller et al., 2006). These signals are not phase locked, unlike MRP. However, the neural generators for both signals are expected to be different. ERD is a decrease in the oscillation frequency while ERS is an increase in the oscillation frequency in the ongoing pattern of the EEG, in response to an induced event. Such EEG signals are processed in the frequency domain and their effects are expressed as relative increment or decrement in power. Furthermore, the effects of ERS and ERD are varied in different frequency ranges. For example, motor events lead to ERD in the low-frequency bands while the same task may lead to ERS in the higher frequency bands. Fig. 2.4 depicts the occurrence of ERS and ERD within a regular EEG signal. The corresponding increase and decrease in relative power can be observed.

2.4 Motor Imagery BCI

BCI can be divided into various categories based on the functionality and the types of the trigger signals. Traditional BCI system can be divided into dependent and independent as well as synchronous and asynchronous. The dependent BCI requires a certain level of motor control from the subjects to operate the BCI system whereas independent BCI does not need any motor control. Another type is synchronous and asynchronous. Synchronous BCI allows the user to operate only during a certain time on the contrary asynchronous BCI can be operated at any time. Furthermore, the BCI can

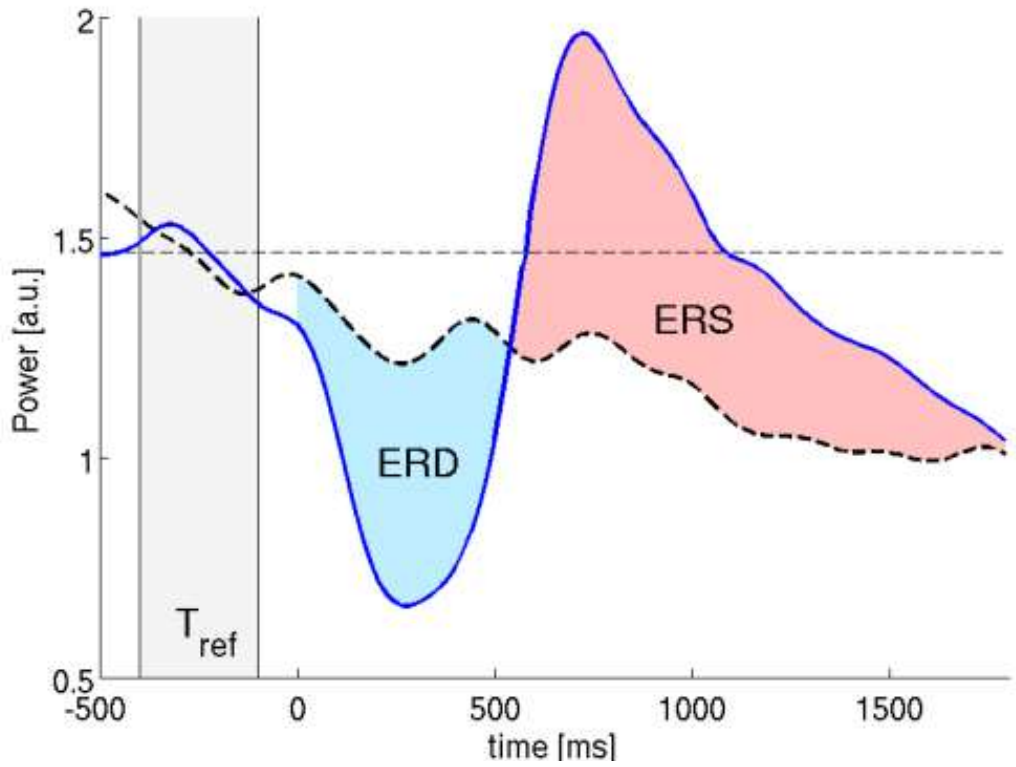


Fig. 2.4 ERS and ERD within a typical EEG signal

Source:http://www.bbci.de

also be divided based on the types of the input trigger brain signals such as Steady State Evoked Potentials (SSEP), P300, Sensorimotor rhythm (SMR), slow cortical potentials and motor imagery. The SSEP and P300 are generated subconsciously when the subject receives the external stimulus. The sensorimotor rhythm and the motor imagery rhythms are generated when the subject performs the actual or imagery movement of the body limbs.

MI-based BCI uses the motor imagery signal which is generated when the users imagine the movement of the body limbs. The steps involved in the MI-based BCI system are summarized in Fig. 2.5. The MI-based BCI is of the synchronous type, where the subject can perform the operation only at a fixed time window. The calibration process starts with the execution of the particular motor imagery movement based on the cue presented in the fixed time window. The subject is asked to repeat this process for several trials. A n number of epoch are extracted and aligned at the start of the cue of each trial from the recorded EEG signals. This extracted epoch of signals is used for further preprocessing. The next step is to select a particular time interval of the acquired signal that represents the ERD/ERS effect. After selecting the particular time interval of the signal, the next step is spectral filtering. The main aim of this is to concentrate on the SMR modulation which is significant in mu and beta rhythm. The filtered signals are

used for the computation of the spatial filter.

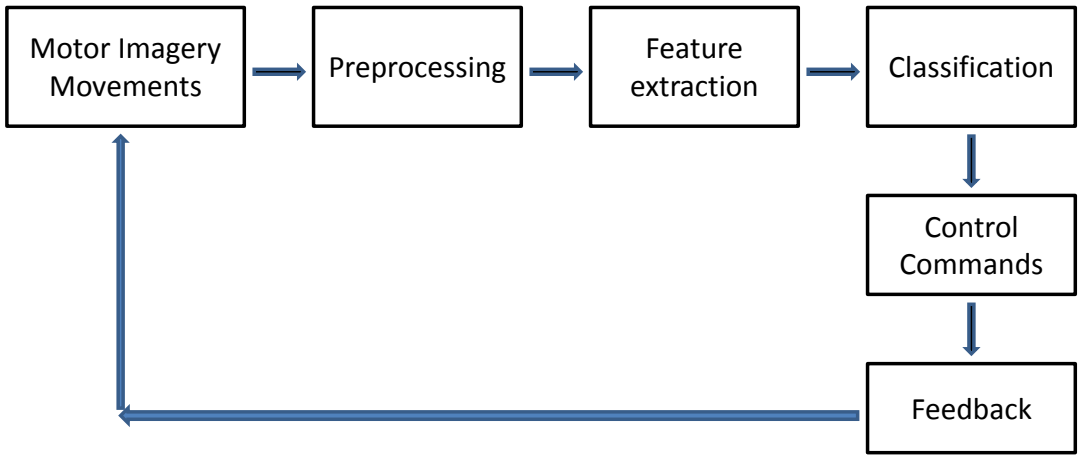


Fig. 2.5 MI based BCI system

Several spatial filtering approaches have been proposed. The commonly used approaches are presented below.

2.4.1 Signal Processing Techniques

- **Common Spatial Pattern (CSP)**(Ramoser et al., 2000; Blankertz et al., 2008): The CSP algorithm computes the spatial filters by maximizing the variance of one class and at the same time minimizing the variance of the other. The obtained spatial filters are used to discriminate the two MI movements. The solution of the CSP is obtained by solving the eigenvalue decomposition. The CSP criterion $J(\mathbf{w})$ can be represented as

$$\min_{\mathbf{w}} \setminus \max_{\mathbf{w}} J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{Cov}_1 \mathbf{w}}{\mathbf{w}^T \mathbf{Cov}_2 \mathbf{w}}, \tag{2.1}$$

  where $\mathbf{Cov_1}$ and $\mathbf{Cov_2}$ are the covariance matrices of class 1 and class 2 and $\mathbf{w}$ is the spatial filter. The equal numbers of the largest and the smallest eigenvalues are selected and the corresponding eigenvectors represent the set of discriminative spatial filters.

- **Independent Component Analysis (ICA)**(Hyvärinen et al., 2004): ICA is one of the most popular Blind Source Separation (BSS) techniques. Let us consider the EEG observation $\mathbf{X}$ which can be denoted as a linear combination of the

independent sources $\mathbf{S}$ and the mixing matrix $\mathbf{A}$

$$\mathbf{X} = \mathbf{AS}. \tag{2.2}$$

The estimated EEG signal $\mathbf{Y}$ is given by

$$\mathbf{Y} = \mathbf{BX}, \tag{2.3}$$

where $\mathbf{B}$ is a unmixing matrix and $\mathbf{B} = \mathbf{A^{-1}}$. The main aim of ICA is to obtain the unmixing matrix using the statistical information of the observation. The observed EEG signals can be considered as mixing source signals from different regions of the brain. ICA can be used for separating the individual source from the mixing model as well as it can be used for removing the artifacts such as eye movements from the acquired EEG signals. Therefore, ICA keeps only the relevant information which enhances the signal-to-noise ratio of the signal.

- **Principal Component Analysis (PCA)**(Hotelling, 1933): It is used for source extraction as well as dimensionality reduction. The main objective of PCA is to perform a linear transformation of the observation into a set of new components which are known as principal components with less dimension. The constrain of the linear transformation is that the first principal component has to have the largest variance followed by the remaining components. The same is followed for the second principal component. This transformation gives the principal components that are uncorrelated with each other. By performing PCA, the input data are projected into a space of eigenvectors. The eigenvectors are computed using the covariance matrix of the input signal $\mathbf{x}(t)$. The covariance matrix $\mathbf{Cov}$ is given by

$$\mathbf{Cov} = \sum_{i=1}^{T} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T, \tag{2.4}$$

where $\bar{\mathbf{x}}$ is the mean which is obtained by

$$\bar{\mathbf{x}} = \tfrac{1}{T} \sum_{i=1}^{T} \mathbf{x}_i. \tag{2.5}$$

The ranking of the eigenvectors is done based on the eigenvalues. The eigenvectors with the highest eigenvalues are taken as the first principal component.

2.4.2 Classification

The next step is to translate the extracted features into control commands. This can be done using classification algorithms. The classifier being used in the field of BCI are

linear classifier (Pfurtscheller, 1999), non-linear classifier (Rezaei et al., 2006), neural networks (Hiraiwa et al., 1990), nearest neighbor classifier (Blankertz et al., 2002) and the combination of classifiers (Pfurtscheller et al., 1993). Among these classifiers, the linear classifier is commonly used for MI discrimination. Two popular linear classifiers for the motor imagery signals are the Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM).

- **Linear Discriminant Analysis (LDA)**:The main aim of LDA (Duda and Hart, 1973) is to project the multidimensional data into a reduced dimensional subspace with higher class separability. LDA approach mainly considers the data for each class as a model of probability density functions. The class of the input data is determined by the larger value of probability density function from the others. LDA assumes that all the classes have a normal distribution and have the same covariance matrix. Let us consider there are c classes and $\mathbf{x} = [x_1, \ldots, x_n]^T$ be the samples to be classified, where n represents the no. of samples. The mean, $\bar{x}$ and the global covariance matrix Cov , can be represented as:

$$\bar{x}_c = \frac{1}{n_k} \sum_{i=1}^{c} x_i \tag{2.6}$$

$$Cov_c = \frac{1}{n_c} \sum_{i=1}^{c} (x_t - \bar{x}_i)(x_t - \bar{x}_i)^T. \tag{2.7}$$

  Then, the classification of data point x is done by

$$g(x) = \arg\max_c x_t {Cov_c}^{-1} \bar{x}_c - \frac{1}{2} \bar{x}_c^T {Cov_c}^{-1} \bar{x}_c \tag{2.8}$$

  which decision boundary is a linear function. The class of x is determined by the objective function given in Eqn. 2.8. The LDA is mainly used for binary classification, but it can also be used for multiclass problem.

- **Support Vector Machine (SVM)**: The SVM (Cortes and Vapnik, 1995) approach has wide applications in the fields of machine learning and pattern recognition. It has the ability to deal with high dimensional and non-linear data. The SVM with kernel can generate non-linear decision boundaries which makes it suitable for discriminating non-linearly separable data. In this method, the data was mapped into a high-dimensional space where the data is spread in such a way that a linear hyper-plane can be fitted. The decision function for kernel-based SVM can be defined by:

$$g(x) = sgn(Cov_{i=1}^{T} \alpha_i c_i k(x, x_i) + b) \tag{2.9}$$

where $x = [1, \ldots, t]$ is the set of training samples, c represents the class labels, $\alpha_i \geq 0$ is a Lagrangian multiplier which is a solution of the quadratic optimization problem, $k(x, x_i)$ represents the kernel and b is the bias. The selection of kernel and setting of the hyperparameters value are important steps in designing SVM classifier. For the experiment in this study, Gaussian RBF is selected which can be defined as

$$k(x, x_i) = e^{-\gamma||x-x_i||^2} \tag{2.10}$$

where $\gamma = \frac{1}{2\sigma^2} > 0$, controls the width of the Gaussian function, $||x - x_i||$ is the norm of x. Moreover, SVM is insensitive to overtraining which makes it suitable for various applications.

2.5 Conclusions

This chapter presents the related background of the brain computer interface system. In the first section, the anatomy and physiology of the human brain involved in motor control were discussed. In the following section, EEG acquisition, EEG rhythms, EEG artifacts and ERS/ERD have been explained. The final section dealt with the steps involved in MI-based BCI. The commonly used signal processing techniques and classifiers used for motor imagery signals classification are also presented in this chapter.

CHAPTER 3

Spatial Filtering Methods

Signal processing techniques mainly aim to de-noise the noisy observed signals and enhance the signal to noise ratio in order to extract the relevant/important information from the signals. Generally, EEG signals are very noisy and easily affected by movement of eyes and muscles. However, it is difficult to remove these artifacts without losing relevant information. Furthermore, it is necessary to filter unrelated brain activity and retain only signals of interest. Therefore preprocessing of the signal is necessary before extracting the required features. The preprocessing of motor imagery EEG signals for BCI applications is mainly done by bandpass filtering the signal together with other filtering techniques. The spatial filtering technique is considered to be quite effective for discrimination of motor imagery EEG signals. The motivation of this chapter is to study the existing spatial filtering and other filtering techniques for this application. The CSP algorithm, regularized CSP algorithms and other variants of CSP are discussed in this chapter.

3.1 Common Spatial Pattern Algorithm

The CSP algorithm was first presented as a feature classification algorithm in (Fukunaga and KoonTz, 1970). Initially, it was used for detection of abnormalities in clinical EEG (Koles, 1991) and also used for the classification between normal and abnormal EEG (Koles et al., 1994). Later, it was used for the discrimination of two class movements from the single trial EEG (Müller-Gerking et al., 1999). A similar study had been performed for discrimination of motor imagery movements from multichannel EEG data.

Let us consider the EEG signals $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_n(t)]^T$ recorded using n channels during the left and right hand MI movements. The main objective of CSP algorithm is to compute spatial filters that discriminate the two MI movements by maximizing the ratio of the variance of signals between the two classes. The CSP objective function can be considered as a Rayleigh quotient maximization problem as:

$$\min_{\mathbf{w}} \setminus \max_{\mathbf{w}} J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{Cov}_1 \mathbf{w}}{\mathbf{w}^T \mathbf{Cov}_2 \mathbf{w}} \tag{3.1}$$

where $\mathbf{w} \in \mathbb{R}^n$ is the spatial filter to be optimized, $\mathbf{Cov}_1$ and $\mathbf{Cov}_2$ denotes the average covariance matrices of class 1 and class 2. The covariance matrix of $\mathbf{x}(t)$ with zero mean for class c is given by

$$\mathbf{Cov}_c = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}(t) - \bar{\mathbf{x}})(\mathbf{x}(t) - \bar{\mathbf{x}})^T \quad c \in \{1, 2\}. \tag{3.2}$$

The solution of Eqn. (3.1) is computed by solving the Generalized Eigenvalue (GEV) decomposition of

$$\mathbf{Cov}_1\mathbf{w}_1 = \lambda\mathbf{Cov}_2\mathbf{w}_1, \tag{3.3}$$

where λ denotes the eigenvalues. The eigenvectors are sorted based on their discriminative abilities. The eigenvector with the largest eigenvalue is considered to have highest discriminative ability than the others. The spatial filters $\mathbf{W}= [\mathbf{w}_1, \ldots, \mathbf{w}_p]$ (where p is the total number of filters selected) for both the class are obtained by selecting $p/2$ eigenvectors with greatest and smallest eigenvalues.

3.2 Divergence Based CSP Approaches

Divergence, the dissimilarity measures between the two distributions, are commonly used in pattern recognition and machine learning techniques. Lately, it has been used for the robust discrimination of motor imagery movements in BCI applications. The solution of the CSP was represented with the optimization of the symmetric Kullback divergence (sKL) in (Wang, 2012; Samek, Blythe, Müller and Kawanabe, 2013; Samek et al., 2014). The sKL divergence ($Div_{sKL}(.||.)$) between the two probability distributions $p(y_i|c_1)$ and $p(y_i|c_2)$ can be represented as

$$Div_{sKL}(p(y_i|c_1)||p(y_i|c_2)) = \int (p(y_i|c_1)\log\frac{(p(y_i|c_1)}{(p(y_i|c_2)}dy_i + \int p(y_i|c_2)\log\frac{p(y_i|c_2)}{p(y_i|c_1)}dy_i, \tag{3.4}$$

where $log(.)$ is the logarithmic function. The sKL divergence ($D_{sKL}(.||.)$) between the class covariance matrices as

$$D_{sKL}(\mathbf{W}^T\mathbf{Cov}_1\mathbf{W}||\mathbf{W}^T\mathbf{Cov}_2\mathbf{W}) = \frac{1}{2}tr((\mathbf{W}^T\mathbf{Cov}_1\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Cov}_2\mathbf{W}) \tag{3.5}$$
$$+ (\mathbf{W}^T\mathbf{Cov}_2\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Cov}_1\mathbf{W})) - 2\mathbf{I}.$$

It is already shown in (Samek et al., 2014) that the subspace of solutions obtained by maximizing the sKL divergence coincides with the subspace of the CSP solutions. The solution of sKL divergence is obtained by

$$\mathbf{W}_{sKL} = \arg\max_{\mathbf{W}} D_{sKL}(\mathbf{W}^\top\mathbf{Cov_1}\mathbf{W}\|\mathbf{W}^\top\mathbf{Cov_2}\mathbf{W}) \tag{3.6}$$

where, $\mathbf{W} = \mathbf{TR}$ can be separated into a whitening matrix $\mathbf{T}$ and orthogonal matrix $\mathbf{R}$. The optimization can be done with respect to $\mathbf{R}$ using

$$\tilde{J}_{sKL}(\mathbf{R}) = \mathbf{D_{sKL}}(\mathbf{I_d}\mathbf{R}\tilde{\mathbf{Cov_1}}\mathbf{R}^\top\mathbf{I_d}\|\mathbf{I_d}\mathbf{R}\tilde{\mathbf{Cov_2}}\mathbf{R}^\top\mathbf{I_d}) \tag{3.7}$$

where, $\mathbf{I}$ is an identity matrix, $\tilde{\mathbf{Cov}}_c$ is obtained by

$$\tilde{\mathbf{Cov}}_c = \mathbf{T}\mathbf{Cov}_c\mathbf{T}^T, c \in \{1, 2\} \quad \text{and} \quad \mathbf{T}(\mathbf{Cov}_1 + \mathbf{Cov}_2)\mathbf{T}^T = \mathbf{I}.$$

It is well known that KL divergence is not robust in the presence of outliers. Hence, the same group used beta divergence to reduce the influence of the outlier trials (Samek, Blythe, Müller and Kawanabe, 2013; Samek et al., 2014). The beta divergence ($Div_\beta(.||.)$) between the two probability distributions is given as

$$\begin{aligned} Div_\beta(p(y_i|c_1)\|p(y_i|c_2)) = \frac{1}{\beta}\int\Bigg(p(y_i|c_1)^\beta \\ - p(y_i|c_2)^\beta\Bigg)p(y_i|c_1)dy_i - \frac{1}{\beta+1}\int\Bigg(p(y_i|c_1)^{\beta+1} - p(y_i|c_2)^{\beta+1}\Bigg)dy_i, \end{aligned} \tag{3.8}$$

and beta divergence coincides with KL divergence when $\beta \to 0$. The objective function can be represented in terms of the covariance matrix and the solution is obtained by maximizing the sum of the beta divergence $D_\beta(.\|.)$ between the trial wise covariance matrix

$$J_\beta(\mathbf{W}) = \sum_{i=1}^{T} D_\beta(\mathbf{W}^\top\mathbf{Cov}_1^i\mathbf{W}\|\mathbf{W}^\top\mathbf{Cov}_2^i\mathbf{W}). \tag{3.9}$$

Furthermore, the regularized divCSP objective function was proposed as

$$J(\mathbf{w}) = (1-\eta)D(\mathbf{w}^T\mathbf{Cov}_1\mathbf{w}||\mathbf{w}^T\mathbf{Cov}_2\mathbf{w}) - \eta\mathbf{P} \tag{3.10}$$

where $\mathbf{P}$ is the penalty term. This work has been further investigated to jointly optimize the robustness and stationarity as

$$\begin{aligned} J(\mathbf{w}) = (1-\eta)\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{T} D_\beta(\mathcal{N}(0, \mathbf{w}^T\mathbf{Cov}_1^i\mathbf{w})||\mathcal{N}(0, \mathbf{w}^T\mathbf{Cov}_2^j\mathbf{w}) \\ -\eta\frac{1}{2T}\sum_{c=1}^{2}\sum_{i=1}^{T} D_\beta(\mathcal{N}(0, \mathbf{w}^T\mathbf{Cov}_c^i\mathbf{w})||\mathcal{N}(0, \mathbf{w}^T\mathbf{Cov}_c\mathbf{w}) \end{aligned} \tag{3.11}$$

where the joint divergence is computed between the i^{th} trial of one class and the j^{th} trial of the other class with the regularization parameter. The regularization parameter is obtained by computing the divergence between individual trials and the overall data distribution of each class. Later, an alternative approach to beta divergence framework

using heavy tail distributions was proposed. But the heavy tail based model does not work well for the outlier trials which are affected differently (Samek and Müller, 2015).

Recently, a group of researchers proposed a divergence based CSP based on Bhattacharya distance and Gamma divergence (Brandl et al., 2015). The Bhattacharyya distance (D_{Bh}) can be defined as

$$D_{Bh}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{T}(ln(|\bar{\mathbf{Cov}}_1^i + \bar{\mathbf{Cov}}_2^i|) - \frac{1}{2}ln|\bar{\mathbf{Cov}}_1^i| - \frac{1}{2}ln|\bar{\mathbf{Cov}}_2^i| - dln(2)), \quad (3.12)$$

and the gamma divergence (D_γ) is defined as

$$\begin{aligned} D_\gamma(\mathbf{w}) = \frac{1}{4\gamma}\sum_{i=1}^{T}(\frac{1}{2}ln(|\gamma\bar{\mathbf{Cov}}_1^i + \bar{\mathbf{Cov}}_2^i|) + \frac{1}{2}ln(|\bar{\mathbf{Cov}}_1^i + \gamma\bar{\mathbf{Cov}}_2^i|) \\ - ln|\bar{\mathbf{Cov}}_1^i| - ln|\bar{\mathbf{Cov}}_2^i| - dln(2)), \end{aligned} \quad (3.13)$$

where $\bar{\mathbf{Cov}}_c^i$ denotes the projected covariance matrix of i^{th} trial and class c. This approach is more robust than the standard CSP with the heavy contaminated data, but the selection of the parameters is the main challenge of this approach.

3.3 The information theoretic feature extraction framework

Information theory plays a key role in the dimensionality reduction step that extracts the relevant subspaces for classification. Inspired by some other papers in machine learning, the authors of (Grosse-Wentrup and Buss, 2008) adopted an information theoretic feature extraction (ITFE) framework based on the idea of selecting those features which are maximally informative about the class labels. In this way, the desired spatial filters are the ones that maximize the mutual information $I(.;.)$ between the output random variable $\mathbf{w}^\top\mathbf{X}$ and the class random variable C, i.e.,

$$\mathbf{w}_* = \arg\max_{\mathbf{w}} I(\mathbf{w}^\top\mathbf{X}\ ;\ C). \quad (3.14)$$

As it was noted in (Grosse-Wentrup and Buss, 2008), this criterion can be also linked with the minimization of an upper-bound on the probability of classification error. Consider the entropy $H(C)$ and a function

$$U(\gamma) = 1 - 2^{-(H(C)-\gamma)}, \quad (3.15)$$

which was used in (Feder and Merhav, 1994) to obtain an upper-bound for the probability of error

$$P_e \leq U(I(C;Y)). \quad (3.16)$$

Since $U(\gamma)$ is an strictly monotonous descending function, the minimization of the upper-bound of P_e is simply obtained through the maximization of the mutual information criterion

$$J_{ITFE}(\mathbf{w}) = I(C; \mathbf{w}^\top \mathbf{X}). \tag{3.17}$$

Although the samples in each class are assumed to be conditionally Gaussian distributed, the evaluation of this criterion also requires to obtain $h(\mathbf{w}^\top \mathbf{X})$, the differential entropy of the output of the spatial filter. This quantity is non-trivial to evaluate, so it was approximated in two steps that assume $\mathbf{w}^\top \mathbf{X}$ is nearly Gaussian distributed. In the first step, the differential entropy is approximated with the help of a truncated version of the Edgeworth expansion for a symmetric density (Jones and Sibson, 1987)

$$h(\mathbf{w}^\top \mathbf{X}) \approx h_g(\mathbf{w}^\top \mathbf{X}) - \tfrac{1}{48} \left(kurt(\mathbf{w}^\top \mathbf{X})\right)^2, \tag{3.18}$$

where $kurt(.)$ denotes the kurtosis and $h_g(\mathbf{w}^\top \mathbf{X})$ denotes the entropy of a Gaussian random variable with power $E[|\mathbf{w}^\top \mathbf{X}|^2]$. The second step consists in approximating this kurtosis by one of the Gaussian random variables with the same power. In this way, the authors of (Grosse-Wentrup and Buss, 2008) arrive at the approximated mutual information criterion that maximizes

$$\begin{aligned} \tilde{J}_{ITFE}(\mathbf{w}) \equiv & -\tfrac{1}{2} \sum_{k=1}^{n_c} P(c_k) \log_2 \left(\mathbf{w}^\top \mathbf{Cov}_k \mathbf{w}\right) \\ & - \tfrac{3}{16} \left(\sum_{k=1}^{n_c} P(c_k) \left((\mathbf{w}^\top \mathbf{Cov}_k \mathbf{w})^2 - 1 \right) \right)^2 \\ \approx & \; J_{ITFE}(\mathbf{w}), \end{aligned} \tag{3.19}$$

where n_c is the number of classes and $\mathbf{Cov_c}$ denotes the conditional covariance matrix of the c^{th} class.

On the one hand, considering only two classes ($n_c = 2$), it is shown that the solution of the ITFE criterion coincides with the solution of CSP. On the other hand, for multiclass scenarios ($n_c > 2$), it is proposed to use a Joint Approximate Diagonalization (JAD) (which we referred as JADE in this thesis) procedure for obtaining the independent sources of the observations and then retain only those sources which maximize the approximated mutual information with the class labels.

3.4 Probabilistic CSP

Overfitting is one of the challenges of CSP algorithm, to address this problem the authors of (Wu et al., 2015) represented the CSP algorithm in probabilistic modelling. The probabilistic model of EEG signals can be represented as

$$\mathbf{X} = \mathbf{AS}, \tag{3.20}$$

where $\mathbf{S} \sim \mathcal{N}(0, \mathbf{Cov}_s)$ and $\mathbf{Cov}_s = diag(\lambda)$. The connection between CSP and the probabilistic model can be defined by

$$\mathbf{W} = \hat{\mathbf{A}}^{-1} \tag{3.21}$$

where $\hat{\mathbf{A}}$ is the maximum likelihood of $\mathbf{A}$. The probabilistic model in the presence of noise is presented as

$$\mathbf{X} = \mathbf{AS} + \mathbf{N}. \tag{3.22}$$

here $\mathbf{A}_n \sim \mathcal{N}(0, \mathbf{Cov})$, $\mathbf{S} \sim \mathcal{N}(0, \mathbf{Cov}_s)$, $\mathbf{E} \sim \mathcal{N}(0, \mathbf{Cov}_n)$. $\mathbf{N}$ is the additive Guassian noise with the covariance matrix $\mathbf{Cov}_n$. Unlike assuming $m = n$, this model assumes $m \leq n$, where m is the number of source and n denotes the EEG channels. The solution of Eqn. (3.22) considers both the spatial and temporal information of the source space. However, this solution is likely to converge to local optima and the determination of $\mathbf{Cov}, \mathbf{Cov}_n$ and $\mathbf{Cov}_s$ are difficult since these parameters are unknown.

Maximum-a-Posteriori CSP (MAP-CSP) tries to address this problem by estimating $\{\mathbf{A}, \mathbf{S}\}$ together in the presence of the isotropic noise

$$\mathbf{X} = \mathbf{AS} + \mathbf{N} \tag{3.23}$$

$$\mathbf{A}_n \sim \mathcal{N}(0, \mathbf{Cov}), \mathbf{S} \sim \mathcal{N}(0, \mathbf{Cov}_s), \mathbf{E} \sim \mathcal{N}(0, \mathbf{Cov}_n\mathbf{I}) \tag{3.24}$$

$$\mathbf{Cov}_s \sim \prod_m \mathcal{G}a^{-1}(\alpha, \beta), \mathbf{Cov}_n \sim \mathcal{G}a^{-1}(\alpha, \beta) \tag{3.25}$$

where,

$$\mathcal{G}a^{-1}(x|\alpha, \beta) := \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} exp(-\beta/x)$$

is the inverse gamma distribution and also assumed that $\alpha \to 0, \beta \to 0$. In contrast to MAP-CSP, Variational Bayesian CSP (VB-CSP) estimated the full posterior distribution by considering a more generalized noise. Following the same model as presented in Eqn. (3.23), VB-CSP algorithm considers

$$\mathbf{A}_n \sim \mathcal{N}(0, \mathbf{Cov}), \mathbf{S} \sim \mathcal{N}(0, \mathbf{Cov}_s), \mathbf{E} \sim \mathcal{N}(0, \mathbf{Cov}_n) \tag{3.26}$$

$$\mathbf{Cov} \sim \prod_m \mathcal{G}a^{-1}(\alpha, \beta), \mathbf{Cov}_s \sim \prod_m \mathcal{G}a^{-1}(\alpha, \beta), \mathbf{Cov}_n \sim \prod_n \mathcal{G}a^{-1}(\alpha, \beta). \tag{3.27}$$

This model can be represented as Bayesian matrix co-factorization model for the signal of the two class (Salakhutdinov and Mnih, 2008).

3.5 Other CSP Variants

As mentioned above, the CSP algorithm computes spatial filters using the covariance matrices. The presence of the outliers leads to poor classification performance. To address this problem, various regularization approaches have been proposed. Regularization is one of the most common approaches used in machine learning for developing a robust system. The regularization of the CSP is mainly done either in the estimation of covariance matrices or by including a penalty term in the objective function. The regularized CSP objective functions can be represented as

$$\tilde{J}_1(\mathbf{w}) = \frac{\mathbf{w}^T \tilde{\mathbf{Cov}}_1 \mathbf{w}}{\mathbf{w}^T \tilde{\mathbf{Cov}}_2 \mathbf{w} + \eta \mathbf{P}(\mathbf{w})}, \tag{3.28}$$

$$\tilde{J}_2(\mathbf{w}) = \frac{\mathbf{w}^T \tilde{\mathbf{Cov}}_2 \mathbf{w}}{\mathbf{w}^T \tilde{\mathbf{Cov}}_1 \mathbf{w} + \eta \mathbf{P}(\mathbf{w})}, \tag{3.29}$$

where $\mathbf{P}$ is the penalty term, η is the regularization parameter and $\tilde{\mathbf{Cov}}_c$ is the estimated covariance matrix of class c. Various approaches are discussed in the following.

The Composite Common Spatial Pattern (cCSP) algorithm (Kang et al., 2009) estimated the covariance matrix by including the information of a subject with similar characteristics. The dissimilarity between subject i and subject j is obtained by computing the KL divergence i.e. $D_{KL}(i,j)$. The estimated covariance matrix can be represented as

$$\begin{aligned} \tilde{\mathbf{Cov}}_c^i &= (1-\eta)\mathbf{Cov}_c^i + \eta \mathbf{G_c} \\ &= (1-\eta)\mathbf{Cov}_c^i + \eta \sum_{j \neq i} \alpha_{ij} \mathbf{Cov}_c^j, \end{aligned} \tag{3.30}$$

where $\mathbf{G}_c = \sum_{j \neq i} \alpha_{ij} \mathbf{Cov}_c^j$ is the generic matrix, $\mathbf{Cov}_c^i$ and $\mathbf{Cov}_c^j$ denotes the data distributions of subject i and subject j for class c, η is the regularizing parameter and α_{ji} denotes the weight of the subjects with similar characteristics which is obtained by

$$\alpha_{ij} = \frac{1}{Z^i} \frac{1}{D_{KL}(i,j)}, \tag{3.31}$$

where,

$$Z^i = \sum_{k \neq i} \frac{1}{D_{KL}(i,k)}, \tag{3.32}$$

is the normalization for subject i. Other approaches include shrinking of covariance matrix towards both the generic and the identity matrix (Lu et al., 2009, 2010), using selected subjects data (Lotte and Guan, 2011) and by using M-estimators (Yong et al., 2008*a*; Kawanabe and Vidaurre, 2009). The stationary subspace analysis (SSA) algorithm (Von Bünau et al., 2009) mainly aims in extracting the stationary sources from the multidimensional EEG signals. The EEG signals can be represented as combination

of the mixing matrix $\mathbf{A}$ and the sources which consists of the stationary $\mathbf{s}^s(t)$ and non stationary $\mathbf{s}^n(t)$ components

$$\mathbf{x}(t) = \mathbf{As(t)} = [\mathbf{A^s A^n}] \begin{bmatrix} \mathbf{s}^s \\ \mathbf{s}^n \end{bmatrix}. \tag{3.33}$$

The goal of SSA is to obtain the estimate separation matrix $\mathbf{B} = \mathbf{RT}$ that separates the stationary sources, here $\mathbf{T}$ is the whitening matrix and $\mathbf{R}$ is an orthogonal matrix. The optimization is performed by minimizing the objective function in each step which is given by

$$J_B(\mathbf{R}) = \sum_{i=1}^{T} D_{KL}[\mathcal{N}(\hat{\mu}_i^s, \hat{\mathbf{Cov}}_i^s)||\mathcal{N}(0, \mathbf{I})], \tag{3.34}$$

where $\hat{\mu}_i^s$ and $\hat{\mathbf{Cov}}_i^s$ is the mean and covariance of the i-th epoch, $\mathcal{N}(\hat{\mu}_i^s, \hat{\mathbf{Cov}}_i^s)$ is the distribution of the stationary sources in each epoch and $\mathcal{N}(0, \mathbf{I})$ represents the normal distribution. The extracted stationary part of the signal is used for the computation of CSP filter (Von Bünau et al., 2010). The extension of this approach known as groupSSA was proposed by Samek et al. (Samek et al., 2011) considering a group of trials and computing the stationary components from each group, which is given by

$$J_B(\mathbf{R}) = \sum_{i=1}^{M} \sum_{j=1}^{N_i} D_{KL}[\mathcal{N}(\hat{\mu}_{ij}^s, \hat{\mathbf{Cov}}_{ij}^s)||\mathcal{N}(\bar{\mu}_j^s, \bar{\mathbf{Cov}}_j^s)] \tag{3.35}$$

where M is the number of groups, N_i is the number of epochs in group i, $\mathcal{N}(\hat{\mu}_{ij}^s, \hat{\mathbf{Cov}}_{ij}^s)$ is the distribution of epoch j in group i and $\mathcal{N}(\bar{\mu}_j^s, \bar{\mathbf{Cov}}_j^s)$ is the average distribution. The same group further extended this approach by including a discriminative term in the groupSSA objective function (Samek, Müller, Kawanabe and Vidaurre, 2012). This method not only considers the stationary components but also considers the discriminative information. The discriminative term is

$$J_B(\mathbf{R}) = D_{KL}[\mathcal{N}(\bar{\mu}_1^s, \bar{\mathbf{Cov}}_1^s)||\mathcal{N}(\bar{\mu}_2^s, \bar{\mathbf{Cov}}_2^s)]. \tag{3.36}$$

This term is subtracted from the objective function given in Eqn. (3.35). After obtaining the stationary and discriminative components, the standard CSP is computed similarly with (Von Bünau et al., 2010).

Another approach of regularizing the CSP algorithm is by including a penalty term in the standard CSP objective function. Unlike the method present in (Grosse-Wentrup et al., 2009), the authors of (Lotte and Guan, 2010*b*) proposed an approach to consider spatial information without any priori information. The penalty term $\mathbf{P}(\mathbf{w})$ is given by

$$\mathbf{P}(\mathbf{w}) = \mathbf{w}^T(\mathbf{D} - \mathbf{G})\mathbf{w}, \tag{3.37}$$

where,

$$\mathbf{G}_{ij} = exp(-\frac{1}{2}\frac{||\mathbf{v}_i - \mathbf{v}_j||^2}{r^2}) \quad \text{and} \quad \mathbf{D}_{ii} = \sum_j \mathbf{G}_{ij}$$

$\mathbf{v}_i$ and $\mathbf{v}_j$ are the vectors containing the co-ordinates of the i^{th} and j^{th} electrodes, r is the closest distance between the two electrodes.

The standard CSP computes the spatial filter based on the variance using l_2 norm which makes the performance of CSP affected with outliers. As EEG signals consist of nonstationarities, the utilization of l_2 norm leads to magnify the effect of noise which hinders the performance. Therefore to reduce the effect of outliers, the authors of (Wang et al., 2012) uses l_1 norm ($||.||_1$) for the computation of CSP filter which is referred as CSP-l_1

$$\tilde{J}_1(\mathbf{w}) = \frac{||\mathbf{w}^T\mathbf{X}||_1}{||\mathbf{w}^T\mathbf{Y}||_1} = \frac{\sum_{i=1}^{T_x}|\mathbf{w}^T\mathbf{x}_i|_1}{\sum_{j=1}^{T_y}|\mathbf{w}^T\mathbf{y}_j|_1}. \tag{3.38}$$

Later, a more generalized optimization function was proposed in (Park and Chung, 2013) using l_p norm ($||.||_p$) instead of using l_1 or l_2 norm, which is given by

$$\tilde{J}_p(\mathbf{w}) = \frac{||\mathbf{w}^T\mathbf{X}||_p}{||\mathbf{w}^T\mathbf{Y}||_p} = \frac{[\sum_{i=1}^{T_x}|\mathbf{w}^T\mathbf{x}_i|^p]^{(1/p)}}{[\sum_{j=1}^{T_y}|\mathbf{w}^T\mathbf{y}_j|^p]^{(1/p)}}. \tag{3.39}$$

Another method of attending the nonstationarities nature of EEG signal is to use a penalty term, computed using the additional measurement like EOG or EMG signal, in the denominator of the CSP objective function (Blankertz et al., 2007). The stationary CSP (sCSP) maximizes the difference between class variance and at the same time keeps it stable within trials. The sCSP was first proposed in (Wojcikiewicz et al., 2011*b*,*a*) and later extended in (Samek, Vidaurre, Müller and Kawanabe, 2012). The penalty term is computed using a chunk of trials

$$\mathbf{P}_{sCSP}^{(k)} = |(\mathbf{Cov}_c^k - \mathbf{Cov}_c)|_+, \tag{3.40}$$

where $|.|_+$ is the non-negative truncation operator that flips the negative eigenvalues. The sCSP method can only extract the stationary feature but it is unable to address the problem of over-fitting. Therefore, the second extension mainly aims at tackling the problem of over-fitting by modifying the objective function as:

$$\tilde{J}(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{Cov}_1\mathbf{w}}{\mathbf{w}^T(\mathbf{Cov}_1 + \mathbf{Cov}_2)\mathbf{w} + \eta_1\mathbf{P}_{sCSP}(\mathbf{w}) + \eta_2\mathbf{P}_{TRCSP}(\mathbf{w})}, \tag{3.41}$$

where $\mathbf{P_{TRCSP}}$ is the Tikhonov regularizer similar with that used in (Lotte and Guan,

2011). Since the sCSP involve flipping of the signs, for some cases it doesn't lead to an optimal solution. Therefore, the authors of (Arvaneh et al., 2011*a*, 2013) proposed a well-defined optimization function

$$\min_{\mathbf{w_i}} \sum_{i=1}^{p/2} \mathbf{w}_i \mathbf{Cov}_2 \mathbf{w}_i^T + \sum_{i=p/2+1}^{p} \mathbf{w}_i \mathbf{Cov}_1 \mathbf{w}_i^T, \tag{3.42}$$

subject to:

$$\mathbf{w}_i(\mathbf{Cov}_1 + \mathbf{Cov}_2)\mathbf{w}_i^T = 1 \quad i = 1, 2, \ldots, p$$
$$\mathbf{w}_i(\mathbf{Cov}_1 + \mathbf{Cov}_2)\mathbf{w}_j^T = 1 \quad i, j = 1, 2, \ldots, p \quad i \neq j$$

This can be reformulated in the presence of the penalty term as

$$\min_{\mathbf{w_i}} (1-\eta) \sum_{i=1}^{p/2} \mathbf{w}_i \mathbf{Cov}_2 \mathbf{w}_i^T + \sum_{i=p/2+1}^{p} \mathbf{w}_i \mathbf{Cov}_1 \mathbf{w}_i^T + \eta \mathbf{P}(\mathbf{w}), \tag{3.43}$$

where,

$$\mathbf{P}(\mathbf{w}) = \frac{1}{2} \sum_{c=1}^{2} \frac{1}{T} \sum_{i=1}^{T} D_{KL}(N(0, \mathbf{w}\mathbf{Cov}_c^i \mathbf{w}^T) || N(0, \mathbf{w}\mathbf{Cov}_c \mathbf{w}^T))$$

subject to:

$$\mathbf{w}_i(\mathbf{Cov}_1 + \mathbf{Cov}_2)\mathbf{w}_i^T = 1 \quad i = 1, 2, \ldots, p$$
$$\mathbf{w}_i(\mathbf{Cov}_1 + \mathbf{Cov}_2)\mathbf{w}_j^T = 1 \quad i, j = 1, 2, \ldots, p \quad i \neq j$$

where the penalty term for each class is the KL divergence between the trial covariance and the average covariance matrices.

The regularization using l_1 norm for estimation of covariance matrix proposed in (Wang et al., 2012). It mainly aims at removing the outliers of large deviation, but noises of small deviation are also present in the EEG signals. The CSP-l_1 approach fails to consider the effect of noise. Therefore, CSP-l_1 had been extended in (Wang and Li, 2016) by modifying the objective function of the CSP-l_1 approach as

$$J(\mathbf{w}) = \frac{\frac{1}{T_x} \sum_{i=1}^{T_x} \sum_{l=1}^{t} |\mathbf{w}^T \mathbf{x}_l^i|}{\frac{1}{T_y} \sum_{j=1}^{T_y} \sum_{l=1}^{t} |\mathbf{w}^T \mathbf{y}_l^j| + \frac{\gamma}{T_z} \sum_{k=1}^{T_z} \sum_{l=1}^{t-1} |\mathbf{w}^T \mathbf{e}_l^k|}, \quad \mathbf{e}_l^k = \mathbf{z}_l^k - \mathbf{z}_{l+1}^k, \tag{3.44}$$

where $\mathbf{z}_l^k$ are the trials used for noise modelling and could be obtained using all the trials of both the classes. $\mathbf{x}_l$ and $\mathbf{y}_l$ are the EEG signals of the two class. T denotes the number of trials and t is the number of sample points during a trial segment.

Another approach for regularizing CSP is by finding sparse spatial filters. The authors of (Arvaneh et al., 2011*a*) obtained the sparse solution by modifying the normalization term to l_1/l_2 norm

$$\min_{\mathbf{w_i}}(1-\eta)\left(\sum_{i=1}^{p/2}\mathbf{w}_i\mathbf{Cov}_2\mathbf{w}_i^T+\sum_{i=p/2+1}^{p}\mathbf{w}_i\mathbf{Cov}_1\mathbf{w}_i^T\right)+\eta\sum_{i=1}^{p}\frac{||w_i||_1}{||w_i||_2}, \tag{3.45}$$

where,

$$\mathbf{P}(\mathbf{w})=\frac{1}{2}\sum_{c=1}^{2}\frac{1}{T}\sum_{i=1}^{T}D(N(0,\mathbf{w}\mathbf{Cov}_c^i\mathbf{w}^T)||N(0,\mathbf{w}\mathbf{Cov}_c\mathbf{w}^T))$$

subject to:

$$\mathbf{w}_i(\mathbf{Cov}_1+\mathbf{Cov}_2)\mathbf{w}_i^T=1 \quad i=1,2,\ldots,p$$
$$\mathbf{w}_i(\mathbf{Cov}_1+\mathbf{Cov}_2)\mathbf{w}_j^T=1 \quad i,j=1,2,\ldots,p \quad i\neq j$$

where the parameter η controls sparseness of the data and accuracy. The solution is obtained by optimizing η. For selecting the channels only two sparse spatial filters are obtained for both the classes. The methods presented in (Yong et al., 2008*b*) and (Arvaneh et al., 2010) are limited to computation of a single spatial filter whereas the method in (Farquhar et al., 2006) can obtain more than one spatial filter but the correlation between the spatial filters are not considered. The previous approaches mainly concentrate in the selection of channels, whereas the authors of (Arvaneh et al., 2011*b*) computed the sparse filter without filtering the channel and also obtained uncorrelated filters.

The other approaches such as Common Spatio-Spectral Pattern (CSSP) (Lemm et al., 2005), Common Sparse Spectral-Spatial Pattern (CSSSP)(Dornhege et al., 2006), Filter Bank CSP (FBCSP) (Ang et al., 2008) that computes both spectral and spatial filters are also proposed in the literature, which is not presented in this thesis.

3.6 Conclusions

In this chapter, the different spatial filtering techniques used for discrimination of motor imagery movements are described. CSP is considered to be the most effective technique for this application. Besides, CSP is easily affected by the presence of the outliers. Hence, several regularized and alternative techniques are proposed which are discussed in this chapter. Moreover, the limitations of existing techniques are also mentioned.

CHAPTER 4

Study of Other CSP Based Approaches

In the previous chapter, the CSP, its variant and other spatial filtering approaches were explained in details. However, the performance of the existing MI based BCI system is still low to use in real time applications. Therefore, different CSP based approaches which include the simplification of CSP solution as a solution of the Rayleigh quotient, the formulation of a discriminative CSP objective function and ICA combined algorithm with CSP are studied and presented in this chapter. Section 4.1 simplifies the CSP objective function as Rayleigh quotient. Section 4.2 presents the discriminative CSP. The ICA corrections to CSP is presented in section 4.3.

4.1 Simplification of the CSP Objective Function

The standard CSP computes the spatial filter by maximizing the ratio of variance between the two classes. As discussed before, the objective function of the CSP algorithm is represented as the Rayleigh quotient

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{Cov}_1 \mathbf{w}}{\mathbf{w}^T \mathbf{Cov}_2 \mathbf{w}}. \tag{4.1}$$

This can also be written as

$$\begin{aligned} \frac{\mathbf{w}^T \mathbf{Cov}_1 \mathbf{w}}{\mathbf{w}^T \mathbf{Cov}_2 \mathbf{w}} &= \frac{\mathbf{w}^T \mathbf{Cov}_2^{\frac{1}{2}} \mathbf{Cov}_2^{-\frac{1}{2}} \mathbf{Cov}_1 \mathbf{Cov}_2^{-\frac{1}{2}} \mathbf{Cov}_2^{\frac{1}{2}} \mathbf{w}}{\mathbf{w}^T \mathbf{Cov}_2^{\frac{1}{2}} \mathbf{Cov}_2^{\frac{1}{2}} \mathbf{w}} \\ &= \frac{\mathbf{w}^T \mathbf{Cov}_2^{\frac{1}{2}} \mathbf{M} \mathbf{Cov}_2^{\frac{1}{2}} \mathbf{w}}{||\mathbf{Cov}_2^{\frac{1}{2}} \mathbf{w}||^2} \\ &= \tilde{\mathbf{w}}^T \mathbf{M} \tilde{\mathbf{w}} \end{aligned} \tag{4.2}$$

where,

$$\mathbf{M} = \mathbf{Cov}_2^{-\frac{1}{2}} \mathbf{Cov}_1 \mathbf{Cov}_2^{-\frac{1}{2}} \tag{4.3}$$

$$\tilde{\mathbf{w}} = \frac{\mathbf{Cov}_2^{\frac{1}{2}} \mathbf{w}}{||\mathbf{Cov}_2^{\frac{1}{2}} \mathbf{w}||} \tag{4.4}$$

The maximization and minimization of Eqn. 4.2 gives the CSP solution.

4.2 Discriminative CSP

The CSP objective function can be reformulated as discriminative CSP which considers the difference between the covariance of the two classes in the numerator. The objective function of the discriminative CSP is given as

$$\min \setminus \max \frac{\mathbf{w}^T(\mathbf{Cov}_1 - \mathbf{Cov}_2)\mathbf{w}}{\mathbf{w}^T(\mathbf{Cov}_1 + \mathbf{Cov}_2)\mathbf{w}}. \tag{4.5}$$

The formulated objective function maximizes the difference of the class covariances from the sum of the class covariance which provides more robustness in the presence of outliers. This solution of the discriminative CSP is equivalent to the solution of the standard CSP which is shown below.

Theorem 4.2.1. *The solution of discriminative CSP is equivalent to the solution of standard CSP*

$$\max \frac{\mathbf{w}^T(\mathbf{Cov}_1 - \mathbf{Cov}_2)\mathbf{w}}{\mathbf{w}^T(\mathbf{Cov}_1 + \mathbf{Cov}_2)\mathbf{w}} = \max \bar{\mathbf{w}}^T \mathbf{Cov}_T^{-\frac{1}{2}}(\mathbf{Cov}_1 - \mathbf{Cov}_2)\mathbf{Cov}_T^{-\frac{1}{2}}\bar{\mathbf{w}}, \tag{4.6}$$

where

$$\bar{\mathbf{w}}^T = \mathbf{w}^T \mathbf{Cov}_T^{\frac{1}{2}}$$

Proof. The equivalence of the solution of CSP and discriminant CSP is shown here

$$\begin{aligned} \frac{\mathbf{w}^T[\mathbf{Cov}_1 - (\mathbf{Cov}_T - \mathbf{Cov}_1)]\mathbf{w}}{\mathbf{w}^T\mathbf{Cov}_T\mathbf{w}} &= \frac{\mathbf{w}^T(2\mathbf{Cov}_1 - \mathbf{Cov}_T)\mathbf{w}}{\mathbf{w}^T\mathbf{Cov}_T\mathbf{w}} \\ &= \bar{\mathbf{w}}^T(2\mathbf{Cov}_T^{-\frac{1}{2}}\mathbf{Cov}_1\mathbf{Cov}_T^{-\frac{1}{2}} - \mathbf{I})\bar{\mathbf{w}} \end{aligned} \tag{4.7}$$

and

$$\mathbf{Cov}_T = (\mathbf{Cov}_1 + \mathbf{Cov}_2).$$

□

The solution of the maximization problem is the same and equal to the generalized eigenvectors of $\mathbf{Cov}_1$ and $\mathbf{Cov}_2$.

4.3 ICA Corrections to CSP

As mentioned in the previous chapter, CSP and ICA are commonly used spatial filtering techniques for discrimination of motor imagery movements. We have tried to incorporate ICA in CSP objective function. It is known that the computation of CSP is easily affected by the presence of the outliers in the data. Therefore, to obtain a more robust objective

function, the ICA techniques are used in combination with the CSP approaches. Let $\mathbf{X}$ be EEG signals which can be represented by the given equation

$$\mathbf{X} = \mathbf{AS}. \tag{4.8}$$

The covariance of $\mathbf{X}$ with zero mean is given by

$$\mathbf{Cov}_x = E[\mathbf{X}\mathbf{X}^T] = \mathbf{A}E[\mathbf{S}\mathbf{S}^T]\mathbf{A}^T. \tag{4.9}$$

Applying whitening transformation $\mathbf{T}$ before the computation of CSP filter

$$\mathbf{T}\mathbf{Cov}_x\mathbf{T}^T = \mathbf{I}, \quad \text{where} \quad \mathbf{T} = \mathbf{Cov}_x^{-\frac{1}{2}} \tag{4.10}$$

and the whitened signal is given by

$$\mathbf{Z} = \mathbf{TAS} = \mathbf{U}^T\mathbf{S}. \tag{4.11}$$

We can further rewrite the above equation as

$$\mathbf{T}^T\mathbf{A}^T\mathbf{Cov}_2^{\frac{-1}{2}}\mathbf{T}^T\mathbf{Cov}_1\mathbf{T}\mathbf{Cov}_2^{\frac{1}{2}}\mathbf{AT} = \mathbf{I}. \tag{4.12}$$

The CSP and ICA objective function can be formulated as

$$\begin{aligned} J(\mathbf{w}) &= \frac{\mathbf{w}^T((1-\eta)\mathbf{Cov}_1 + \eta\mathbf{Cov_P})\mathbf{w}}{\mathbf{w}^T\mathbf{Cov}_T\mathbf{w}} \\ &= \frac{\mathbf{w}^T((1-\eta)\mathbf{Cov}_1 + \eta\mathbf{Cov_P})\mathbf{w}}{\mathbf{w}^T(\mathbf{Cov}_1 + \mathbf{Cov}_2)\mathbf{w}}, \end{aligned} \tag{4.13}$$

where, the penalty term $\mathbf{Cov_P}$ can be the reference covariance computed from the orthogonal subspace to the artifact signals and

$$J(\mathbf{w}) = \begin{cases} CSP & \text{if} \quad \eta = 0, \\ ICA & \text{if} \quad \eta = 1, . \end{cases} \tag{4.14}$$

The solution to this problem is given by

$$\mathbf{Y} = \mathbf{W}^T\mathbf{X} \tag{4.15}$$

where, $\mathbf{W} = [w_1, w_2, \ldots, w_p]$ is the discriminative spatial filters.

4.4 Experimental dataset and study

This discriminative CSP and ICA corrections to CSP approaches have been tested using the publicly available BCI competition IV dataset 2a (Brunner et al., 2008). The dataset

consists of left hand, right hand, foot and tongue motor imagery movements acquired using 22 electrodes from nine healthy subjects. We have considered only the left and right hand motor imagery movements for this study. The MI EEG signals are band-pass filtered between $7-30Hz$ and 10- fold cross validation was applied to obtain the training set and the testing set.

In the first experiment, we have done a performance comparison between the discriminative CSP and the standard CSP. For the second approach, we have used SOBI algorithm to find the related independent components. In this approach, we have introduced a parameter η to select between the CSP and ICA algorithm. Therefore, when $\eta = 0$ then the algorithm is considered as a standard CSP and when $\eta = 1$ then its ICA. If $\eta = 0.5$ then we considered the combination of ICA with CSP algorithm.

4.5 Results

The performance comparison between the discriminative CSP and standard CSP is shown in Table 4.1. From the result, we can observed that both the algorithms give the same performance. The performance results for ICA corrections to CSP approaches for different value of η are shown in Table 4.2. From the observation, we conclude that the standard CSP algorithm (i.e. $\eta = 0$) gives the maximum performance.

Table 4.1 Performance comparison between CSP and discriminative CSP

Methods	Average Accuracy(%)
CSP	81
discriminative CSP	81

Table 4.2 Performance comparison of ICA corrections to CSP for different values of $\eta = [0, 0.5, 1]$

η	Average Accuracy(%)
$\eta = 0$, i.e. CSP	81
$\eta = 0.5$	80
$\eta = 1$ i.e. ICA	67

4.6 Conclusions

This chapter presents different approaches based on CSP. Although the proposed algorithms failed to improve the accuracy, it was interesting to study these. We hope that these approaches can further be developed and modified for obtaining a robust method. Performance results of these studies will not be presented in this thesis.

CHAPTER 5

Optimization of Thin Independent Component Analysis-Common Spatial Pattern (ThinICA-CSP) Algorithm

EEG signals are represented in a very high dimensional space and it is difficult to classify the data present in the high dimensional space. Therefore, the process of reducing the number of dimensions and simultaneously keeping the useful information for discrimination of different MI movements is a necessary step in EEG signal processing for BCI applications. Moreover, the current MI-BCI system has a long way to go before being used for real time applications. One of the approaches to overcome this gap is to increase the number of MI movements (instead of using only two classes). However, increasing the number of MI movements will lead to decrease in the classification performance. Here arises a need for a robust algorithm for classification of multiclass movements.

A criterion to address the problem of discrimination of multiclass MI movements is proposed in this chapter. The chapter is organized as follows. In section 5.1, the commonly used BSS techniques are described. The related works are presented in 5.2. The experimental design which includes the dataset, the pre-processing steps, the proposed ThinICA-CSP criterion and the classifiers used are described in section 5.3. Section 5.4 presents the results obtained.

5.1 Blind Source Separation Techniques and Its Background

The BSS is a well-known signal processing technique commonly used for solving the cocktail party problem. In this problem, multiple numbers of speakers are present in a room and the solution is to obtain the speech signal of the individual speaker. The basic model of BSS problem is shown in Fig. 5.1. To represent it mathematically, let us consider m number of sources and the signal from each source is represented by $\mathbf{s}(t) = s_1(t), s_2(t), \ldots, s_m(t)$. Let n be the number of sensors used for collecting the sources signals and the observed signals at particular time t can be given as $\mathbf{x}(t) = x_1(t), x_2(t), \ldots, x_n(t)$. The observations can be represented as a linear combination of

the source signals as

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1m} \\ a_{21} & \dots & a_{2m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_m(t) \end{bmatrix} \tag{5.1}$$

where $a_{11}, a_{12}, \dots, a_{nm}$ are the unknown weights. The above equation can be expressed as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{5.2}$$

where $\mathbf{A}$ is the unknown mixing matrix. For simplicity, it is assumed that the number of sensors is equal to the number of sources $(n = m)$. The main objective of BSS is to obtain the estimated sources signals $\mathbf{y}(t)$ from the observations $\mathbf{x}(t)$ by using some statistical properties of the sources. The solution is obtained by designing an unmixing matrix $\mathbf{B} \in \mathbb{R}^{n \times m}$ which represents the estimated sources $\mathbf{y}(t)$ when multiplied to the observed signals

$$\mathbf{y}(t) = \mathbf{B}^H \mathbf{x}(t) \tag{5.3}$$

where the unmixing matrix $\mathbf{B}$ is the inverse of the mixing matrix $\mathbf{A}$

$$\mathbf{B}^H = \mathbf{A}^{-1}. \tag{5.4}$$

Blind Source Extraction (BSE) is another approach of source separation where only a single or a few interesting signals are extracted from a large number of observations. In general, BSE is more flexible than BSS as it enables to extract only the interesting subsets of an independent component, unlike BSS that extracts the whole sources. It also reduces the computational complexity. The solution of the BSS can be obtained by extracting the components sequentially or by performing a deflation after each extraction. Alternatively, it can also be solved by using JAD techniques. The PCA and ICA are the commonly used methods for source separation. Although, only the later one can guarantee the identifying ability of the sources (provided that they are independent and non-Guassian)

5.1.1 Principal Component Analysis (PCA)

PCA is one of the commonly used techniques for reducing the dimensional of the multi-dimensional data. It transforms a set of correlated variable to a new set of uncorellated variables which are known as principal components. The term principal component was first introduced by Hotelling (Hotelling, 1933). In PCA transformation, the signals are

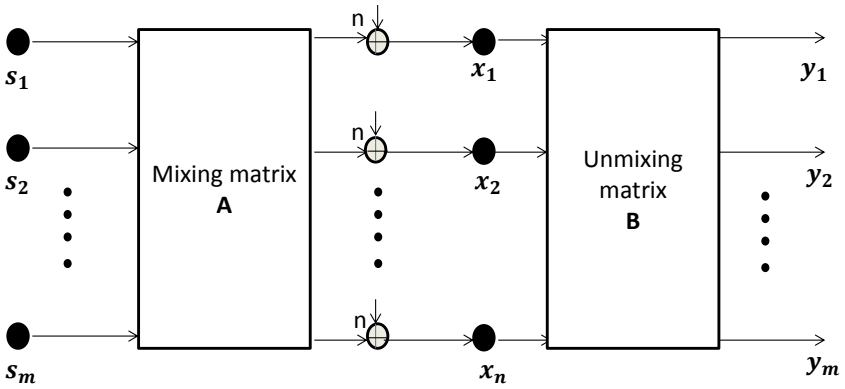


Fig. 5.1 Blind Source Separation. In the figure $\{\mathbf{s_1}, \ldots, \mathbf{s_m}\}$ are the sources, n is the added noise, $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ are the observations,$\{\mathbf{y_1}, \ldots, \mathbf{y_m}\}$ are the principal components.

first centered by subtracting its mean

$$\mathbf{x}(t) \leftarrow \mathbf{x} - E\{\mathbf{x}(t)\}. \tag{5.5}$$

The objective of PCA is to obtain the new set of uncorrelated variables $\mathbf{s(t)}$ with less dimension than the observations $\mathbf{x(t)}$. The computation of PCA is only based on the second order statistics. Whitening is the commonly used approaches in the transformation of signals. The whitening transform first decorrelates the data and later performs the scaling of variance to unit variance. Hence, the covariance of the transformed data is an identity matrix. The whitening matrix $\mathbf{T}$ is computed by

$$\mathbf{T} = \Delta^{-1/2}\mathbf{U}^T \tag{5.6}$$

where $\mathbf{\Delta}$ is the matrix containing eigenvalues of the covariance matrix $\mathbf{Cov}_x$ and the columns of the unitary matrix $\mathbf{U}$ represents the eigenvectors. The whitened signal $\mathbf{z}(t)$ is obtained by multiplying the observations $\mathbf{x}(t)$ by the whitening transform $\mathbf{T}$

$$\mathbf{z}(t) = \mathbf{T}\mathbf{x}(t) \tag{5.7}$$

and

$$\mathbf{Cov}_z = E\{\mathbf{z}(t)\mathbf{z}(t)^T\} \approx \mathbf{I} \in \mathbb{R}^{n \times m}. \tag{5.8}$$

The goal of PCA is to find the matrix $\mathbf{T}$ that minimizes the cross correlation between the whitened signal $\mathbf{z}$.

5.1.2 Independent Component Analysis (ICA)

PCA performs decorrelation of data by using only the second order and considering minimum mean square error whereas ICA uses both the second and higher order statistic for the computation of independent components. Moreover, the solution of PCA is a set of source signals whereas ICA can separate independent components from the set of the extracted sources. ICA has been used in various applications for removing artifacts from EEG signals and separation of brain rhythms (Jung et al., 2000; Makeig et al., 2004). ICA is a successful technique among the various sub-field of BSS that assumes the mutual independence of the sources. ICA linearly decomposes the multichannel observations into independent signals.The extraction model for ICA is similar to that of BSS problem. Let $\mathbf{s(t)}$ be the source signal and $\mathbf{x(t)}$ be the mixture observations

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t). \tag{5.9}$$

Whitening transformation is a step used in ICA algorithm. As described above, it decorrelates the data as well as forces the data to have unit variance. The whitened data $\mathbf{z}(t)$ is given by

$$\mathbf{z}(t) = \mathbf{T}\mathbf{x}(t), \tag{5.10}$$

where $\mathbf{T}$ is the whitening matrix. The main goal of ICA is to obtain the unmixing matrix $\mathbf{B}$, assuming that the sources are independent of each other, the observations are formed by the linear combination of the sources signals, the independent components have non-Gaussian distributions, the number of sources is equal to the number of observations and the mixing matrix $\mathbf{A}$ is invertible. Therefore, the estimated source signal can be obtained by

$$\mathbf{y}(t) = \mathbf{B}^H\mathbf{z}(t), \tag{5.11}$$

where, $\mathbf{B^H} = \mathbf{A^{-1}}$. The extraction method of ICA is shown in Fig. 5.2.

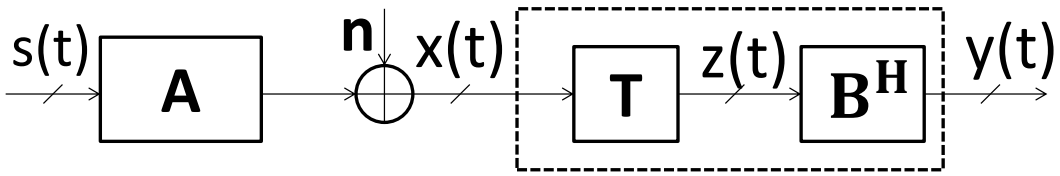


Fig. 5.2 Extraction method for ICA. In the figure $\mathbf{s(t)}$ is the source, $\mathbf{A}$ is the mixing matrix, $\mathbf{n}$ is the added noise, $\mathbf{x(t)}$ is the observation, $\mathbf{T}$ is the whitening matrix, $\mathbf{z(t)}$ is the whitened data, $\mathbf{B}$ is the unmixing matrix, $\mathbf{y(t)}$ is the independent component.

Various ICA techniques have been developed for estimating the mixing matrix such as Information Maximization (InfoMax), Fast Independent Component Analysis (FastICA), Second Order Blind Identification (SOBI) etc. ICA algorithm can be divided into various

types based on the different criteria such as minimization of the mutual information, maximization of the non-Gaussianity and based on the higher order statistic.

- InfoMax algorithm: InfoMax ICA algorithm is based on the minimization of mutual information. The mutual information between the two variables X and Y is defined by

$$I(X,Y) = H(X) - H(X|Y), \tag{5.12}$$

where $H(X)$ is the entropy of variable X and $H(X|Y)$ is the conditional entropy of X when given Y. The conditional entropy $H(X|Y)$ can be obtained by

$$H(X|Y) = H(X,Y) - H(Y), \tag{5.13}$$

where $H(Y)$ is the entropy of variable Y and $H(X,Y)$ is the joint entropy of variable X and Y. Therefore, the entropy of variable X is given as

$$H(X) = -\sum_{i} P(x_i) log P(x_i), \tag{5.14}$$

and the joint entropy of X and Y is

$$H(X,Y) = -\sum_{i,j} P(x_i, y_i) log P(x_i, y_i), \tag{5.15}$$

where $P(x_i)$ is the probability of x_i. InfoMax ICA was first proposed by (Bell and Sejnowski, 1995) in 1995. Later it was proposed by Amari et al. in (Amari et al., 1996) where the unmixing matrix $\mathbf{B}$ is obtained by optimizing the constraint

$$\mathbf{B}(t+1) = \mathbf{B}(t) + \mu(t)(\mathbf{I} - f(\mathbf{B})\mathbf{B}^T)\mathbf{B}(t), \tag{5.16}$$

where t is the step, $\mu(t)$ is the function that specifies the step size, $f(Y)$ depends on the distribution of the sources and sometimes is set equal to $f(Y) = 1 + exp^{-Y^{-1}}$ or $f(Y) = tanh(Y)$.

- SOBI(Belouchrani et al., 1997): SOBI is based on the second order cumulants. This method mainly considers the temporal information from the time-lagged covariance matrices. The sources are obtained by diagonalizing these covariance matrices. The correlation matrix of the signal $\mathbf{R}_x$ is given by

$$\mathbf{R}_x(\tau) \equiv \langle \mathbf{x}(t+\tau)\mathbf{x}^T(t) \rangle = \mathbf{A}\mathbf{R}_s(\tau)\mathbf{A}^T + \delta(\tau)\sigma\mathbf{I}, \tag{5.17}$$

where $\mathbf{R}_s(\tau)$ is the correlation matrix of the source signals with small time delay

τ. SOBI starts with the computation of the whitening matrix $\mathbf{T}$ and the whitened signal is given by

$$\langle \mathbf{T}\mathbf{y}(t)\mathbf{y}(t)^T\mathbf{T}^T\rangle = \mathbf{T}\mathbf{R}_s(0)\mathbf{T}^T = \mathbf{T}\mathbf{A}\mathbf{A}^T\mathbf{T}^T = \mathbf{I}, \tag{5.18}$$

where $\mathbf{TA}$ is a unitary matrix, i.e. $\mathbf{TA} = \mathbf{U}$. Therefore, the mixing matrix $\mathbf{A}$ is equal to

$$\mathbf{A} = \mathbf{T}^{-1}\mathbf{U}. \tag{5.19}$$

- Joint Approximation Diagonalization of Eigenmatrices (JADE): Unlike PCA and SOBI, JADE utilizes the fourth order cumulants for the computation. It tries to minimize the mutual information contained in the cumulant matrices. This is done by obtaining a rotational matrix that diagonalized the cumulant matrices. Cumulants are higher order correlation, therefore it can be considered as a measure of independence. The fourth order cumulants can be represented as:

$$\begin{aligned} cum(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l) = {} & E\{\mathbf{x}_i\mathbf{x}_j\mathbf{x}_k\mathbf{x}_l\} - E\{\mathbf{x}_i\mathbf{x}_j\}E\{\mathbf{x}_k\mathbf{x}_l\} \\ & - E\{\mathbf{x}_i\mathbf{x}_k\}E\{\mathbf{x}_j\mathbf{x}_l\} - E\{\mathbf{x}_i\mathbf{x}_l\}E\{\mathbf{x}_j\mathbf{x}_k\}, \end{aligned} \tag{5.20}$$

where $cum(.)$ represents the cumulants, $\mathbf{x}_{(i,j,k,l)}$ represent the observed signals and $i, j, k, l = 1, \ldots, n$ where n is the number of measured mixtures. The cumulants tensor can be represented in a matrix form by linear transformation

$$F_{ij}(\mathbf{M}) = \sum_{kl} m_{kl} cum(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k, \mathbf{x}_l), \tag{5.21}$$

where m_{kl} is the coefficient. The constraint for JADE is given by

$$J_{JADE}(\Theta) = \sum_i ||diag(\Theta\mathbf{F}(\mathbf{M})_i\Theta^T)||^2, \tag{5.22}$$

where $||diag(.)||^2$ denotes the sum of squares of diagonal elements.

- FastICA (Bingham and Hyvärinen, 2000): In this method, the estimation of source signals is done by enforcing the non-Gaussianity. FastICA algorithm maximizes the non-Gaussianity by using kurtosis. The kurtosis for the random variable after centering is given by

$$kurt(y) = E\{\mathbf{y}^4\} - 3(E\{\mathbf{y}^2\})^2, \quad \text{where} \quad \mathbf{y} = b^T\mathbf{x}. \tag{5.23}$$

Kurtosis is zero for Gaussian distribution. The main objective of FastICA is to

find the weight vectors b using the objective function

$$J(b) = E(b^T\mathbf{x})^4 - 3||w||^4 + \mathbf{P}(||w||^2). \tag{5.24}$$

The constraint $||w|| = 1$ is taken into consideration. $\mathbf{P}$ is a penalty term and several forms of penalty term were proposed in (Hyvärinen and Oja, 1996).

5.2 Related Work

As mentioned earlier, the popular algorithms to attend this issue are the CSP algorithm and various ICA techniques. CSP algorithm is a supervised process whereas ICA is an unsupervised technique where the class labels are unknown. CSP algorithm (Blankertz et al., 2008), which is considered to be the most effective method for discrimination of two class MI movements, is further extended to discriminate four MI movements in (Dornhege et al., 2004). Multiclass CSP considers the multiclass problem as a combination of a binary problem. ICA is commonly used for removing the artifacts from the EEG signals (Jung et al., 2000) and also used for discriminating the *mu* rhythms generated from the sensory motor cortex (Makeig et al., 2004). The authors of (Naeem et al., 2006) and (Brunner et al., 2007) evaluated different ICA algorithms like Infomax, FastICA, SOBI for separation of four-class MI movements. Both the studies conclude that the Infomax performs better than FastICA and SOBI but the performance of Infomax is comparable with the multiclass CSP. Another author proposed a new feature extraction method based on the Infomax algorithm using patterns from the independent components (Zhou et al., 2014).

There are various other different approaches proposed for the classification of multiclass EEG signals. Similar with the multiclass CSP, the multiclass FBCSP (Chin et al., 2009) is an extension of the FBCSP method that uses different frequency bands for discrimination of MI movements. Several other approaches have been proposed for the classification of multiclass MI signals. The other approach for multiclass CSP extension is done by using simultaneous JAD techniques (Dornhege et al., 2004), which were further extended by incorporating information theory for extracting the features (Grosse-Wentrup and Buss, 2008). This approach showed the relation between JAD and BSS. It was observed that the JAD based multiclass approach outperforms the one versus rest CSP approach. The connections between JAD and BSS was demonstrated by forming a new formulation based on maximum likelihood framework (Gouy-Pailler et al., 2010). More recently, a group (Barachant et al., 2012) has proposed the separation of multiclass MI signals by exploiting the Riemannian geometry of the manifold of covariance matrices and using the tangent space for the classification of the features. Other authors proposed a new criterion for multiclass problem based on maximizing

the harmonic mean of the KL divergence between the class covariance matrices (Wang, 2012). The authors of (Nguyen et al., 2012) applied stationary CSP approach together with JAD to solve the multiclass problem. Other approach includes utilizing Bayesian learning method for multitask classification (Zhang et al., 2013). The authors of (Xu et al., 2011; Llera et al., 2014) and (Nicolas-Alonso et al., 2015) used the adaptive processing in discriminating multiclass problem. Another different approaches in solving the multiclass classification is by performing Fourier transformation (Townsend et al., 2006; Ge et al., 2014). This chapter provides the overview of the BSS techniques. A new criterion for the classification of four class MI movements was proposed based on BSS.

5.2.1 Multiclass CSP

The standard CSP method derives the spatial filter by maximizing the variance of one class and simultaneously minimizing the variance of the other class. It is obtained by solving the Rayleigh quotient using the generalized eigenvalue problem

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{Cov}_1 \mathbf{w}}{\mathbf{w}^T \mathbf{Cov}_2 \mathbf{w}}, \tag{5.25}$$

where $\mathbf{Cov}_1$ and $\mathbf{Cov}_2$ denotes the covariance matrices of class 1 and class 2. The resulting eigenvalues are sorted in a descending order and the largest and smallest p no. of features are selected. The corresponding eigenvectors of the selected eigenvalues represent the spatial filters. This method can discriminate only two motor imagery classes.

Therefore, the same concept had been used to extend for discrimination of multiclass movements (Dornhege et al., 2004). In multiclass CSP, the multiclass problem is considered as a binary problem and spatial filters are obtained between one class upon the rest of the class. The eigenvalues were sorted in descending order and the p number of smallest eigenvalues were selected for each class.

5.3 Implementation of the Discrimination of the MI-EEG Signals

The experimental design of the MI based BCI system consists of the following stages: (i) First, EEG signals acquired during the execution of mental tasks were used from BCI competition dataset. (ii) These signals were filtered using a bandpass filter (between particular cut off frequencies) and a small time segment of EEG signals was selected. (iii)The filtered EEG signals were used to determine the spatial filters of the corresponding MI tasks. (iv) The EEG signals were filtered using the obtained spatial filters. (v) The logarithmic variances of the spatially filtered signals were used as features for training the classifiers e.g. LDA and SVM. The overall experimental design of this study is describe briefly in the following sections.

5.3.1 Experimental Dataset

In this study, dataset 2a from BCI competition IV (Brunner et al., 2008) was used. The data was acquired from nine subjects while performing the MI movement of four different task i.e. left hand, right hand, feet and tongue. The dataset consists of two sessions and each session consists of 288 trials. The EEG signals were acquired using 22 electrodes. The acquired signals were sampled at a sampling frequency of 250 Hz and filtered with a pass band filter of cut off frequencies 0.5 Hz and 100 Hz. A notch filter of 50 Hz was used to remove the line noise.

5.3.2 Preprocessing

The *mu* rhythm represents the ERS and ERD of the sensory motor cortex during the MI movements. In order to select the appropriate information, the acquired EEG signals were filtered using a fifth order Butterworth filter with a cut off frequency between 8-30 Hz for each subject. From the filtered signal, the time segment of 0.5-2.5s after the cue has been selected for each trial. For later convenience, the selected signals were concatenated in a classwise order.

5.3.3 The Thin ICA-CSP Criterion and its Implementation

An extension of the ThinICA algorithm, which was proposed in (Cruces, Cichocki and De Lathauwer, 2004) is been proposed in this section. The second and higher order statistics of the observations are combined to obtain a contrast function that we adapt and specialize for EEG processing.

The acquired EEG signals with n recording channels at time t are denoted as $\mathbf{x}(t) \in \mathbb{R}^n$. Their linear model in term of the latent sources is given by

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{5.26}$$

where $\mathbf{A} \in \mathbb{R}^{n \times m}$ is an unknown mixing matrix, $\mathbf{s}(t) \in \mathbb{R}^m$ is the vector of independent sources with covariance matrix equal to the identity $\mathbf{Cov_s} = \mathbf{I}$, while $\mathbf{n}(t) \in \mathbb{R}^m$ denotes additive Gaussian noise of zero mean and covariance matrix $\mathbf{Cov_n} = \sigma_n^2\mathbf{I}$. The prewhitening of the observations is a useful preprocessing that simplifies the ICA problem. The covariance of the observations is

$$\mathbf{Cov_x} = \mathbf{A}\mathbf{A}^T + \sigma_n^2\mathbf{I}. \tag{5.27}$$

Therefore, pre-multiplying the observations with the whitening matrix

$$\mathbf{T} = (\mathbf{Cov}_x - \sigma_n^2\mathbf{I})^{\frac{1}{2}}$$

retains and whitens the principal signal subspace of the observed signals. We denote the prewhitened observations with

$$\begin{aligned} \mathbf{z}(t) &= \mathbf{T}\mathbf{x}(t) && (5.28) \\ &= \mathbf{U}\mathbf{s}(t) + \mathbf{T}\mathbf{n}(t) \in \mathbb{R}^n, && (5.29) \end{aligned}$$

where the orthogonal matrix $\mathbf{U} = \mathbf{TA}$ denotes the residual mixing matrix after prewhitenning.

The simultaneous estimation of $p \leq n$ source signals (plus a non-separable noise component) can be obtained by multiplying the whitened observations by the transpose of the semi-orthogonal matrix $\hat{\mathbf{U}} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_p] \in \mathbb{R}^{n \times p}$. Thus, the outputs or estimated sources can be represented as

$$\mathbf{y}(t) = \hat{\mathbf{U}}^T \mathbf{z}(t). \tag{5.30}$$

Classical ICA algorithms, like (Cruces, Cichocki and Amari, 2004), recover p independent components by maximizing the contrast function based on the sum of the squared kurtosis of the outputs

$$\max \sum_{i=1}^{p} \left| Cum(\mathbf{y}_i(t), \mathbf{y}_i(t), \mathbf{y}_i(t), \mathbf{y}_i(t)) \right|^2, \tag{5.31}$$

over the Stiefel manifold, i.e., subject to the constraint. The function is given by: $\hat{\mathbf{U}}^T \hat{\mathbf{U}} = \mathbf{I}$. Note that, for later convenience, the kurtosis of $\mathbf{y}_i(t)$ is denoted with the fourth-order cumulant $Cum(\mathbf{y}_i(t), \mathbf{y}_i(t), \mathbf{y}_i(t), \mathbf{y}_i(t))$. When the sources are non-stationary, it is convenient to exploit this diversity by splitting the data into K blocks with center at $t_k, k = 1, \ldots, K$. The marginal contrast functions can be evaluated at each split of the data and we can simultaneously maximize the accumulated sum of the marginal contrasts.

$$\begin{aligned} \max \quad & \sum_{k=1}^{K} \sum_{i=1}^{p} \left| Cum(\mathbf{y}_i(t_k), \mathbf{y}_i(t_k), \mathbf{y}_i(t_k), \mathbf{y}_i(t_k)) \right|^2 \\ \text{subject to} \quad & \hat{\mathbf{U}}^T \hat{\mathbf{U}} = \mathbf{I}. \end{aligned} \tag{5.32}$$

However, it is well known that, generally, the variance of the estimates increases with the order of the statistics. Therefore, for short data records, the variance is high and the previous approach is not very precise. A more reliable method consists in allowing the combination of several low-order statistics of the outputs to build the contrast that estimates the demixing matrix. The third order cumulants are usually not considered because they vanish for symmetric distributions, so it seems reasonable to consider only the combination of second and fourth-order statistics of the outputs. Additionally, when

the sources are correlated this additional diversity can be also exploited by combining the second-order cross-cumulants at different delays $\tau \in \mathcal{T}$. All these considerations lead to the proposal of the maximization of the contrast function Ψ_Θ (given below) in the Stiefel manifold

$$\begin{aligned}\Psi_\Theta(\hat{\mathbf{U}}) = \quad & \gamma_4 \sum_{k=1}^{K}\sum_{i=1}^{p} |Cum(y_i(t_k),\ldots,y_i(t_k))|^2 \\ & + \gamma_2 \sum_{k=1}^{K}\sum_{i=1}^{p}\sum_{\tau\in\mathcal{T}} |Cum\left(y_i(t_k+\tau), y_i(t_k)\right)|^2 .\end{aligned} \tag{5.33}$$

Here, γ_2 and γ_4 are weighting proportional to the precision of the respective estimates of the second and higher order statistics. However, the drawback of this contrast function is that it is highly non-linear in $\hat{\mathbf{U}}$, and therefore, it is difficult to optimize.

Fortunately, the previous contrast function can be generalized by decoupling the extraction candidates that appear at the arguments of the cumulants to obtain a new contrast function which is quadratic in each decoupled extraction matrix. The idea is to use four different estimates of the sources $y_i^{[q]}(t_k) = (\hat{\mathbf{U}}^{[q]})^T\mathbf{z}(t_k),\ q = 1,\ldots,4$, to build following the ThinICA-CSP contrast function Φ:

$$\begin{aligned}\Phi(\hat{\mathbf{U}}^{[1]},\ldots,\hat{\mathbf{U}}^{[4]}) \quad = \quad & \\ & \gamma_4 \sum_{k=1}^{K}\sum_{i=1}^{p}\sum_{\tau\in\mathcal{T}} \left|Cum(y_i^{[1]}(t_k),\ldots,y_i^{[4]}(t_k))\right|^2 \\ & + \frac{\gamma_2}{3} \sum_{k=1}^{K}\sum_{i=1}^{p}\sum_{\tau\in\mathcal{T}} \left|Cum\left(y_i^{[1]}(t_k+\tau), y_i^{[2]}(t_k)\right)\right|^2 \\ & + \frac{\gamma_2}{3} \sum_{k=1}^{K}\sum_{i=1}^{p}\sum_{\tau\in\mathcal{T}} \left|Cum\left(y_i^{[1]}(t_k+\tau), y_i^{[3]}(t_k)\right)\right|^2 \\ & + \frac{\gamma_2}{3} \sum_{k=1}^{K}\sum_{i=1}^{p}\sum_{\tau\in\mathcal{T}} \left|Cum\left(y_i^{[2]}(t_k+\tau), y_i^{[3]}(t_k)\right)\right|^2\end{aligned} \tag{5.34}$$

which is quadratic with respect to each $\hat{\mathbf{U}}^{[q]}$ and should be independently maximized for each extraction candidate $q = 1,\ldots,4$ under the constraint $(\hat{\mathbf{U}}^{[q]})^T\hat{\mathbf{U}}^{[q]} = \mathbf{I}$. At its maximum value, all the estimates $y_i^{[1]}(t) = \ldots = y_i^{[4]}(t)$ will agree and recover p independent sources. We omit here the details of the proof of the contrast function nature of Eqn. 5.34 but the required steps are similar to those presented in (Cruces and Cichocki, 2003).

The ThinICA-CSP contrast function in Eqn. 5.34 is sequentially maximized with respect to each extraction candidate $\hat{\mathbf{U}}^{[q]},\ q = 1,\ldots,4$. For this purpose its gradient $\nabla_{\hat{\mathbf{U}}^{[q]}}\Phi$ is evaluated and later, with the help of its singular value decomposition, this

gradient is orthogonally projected onto the Stiefel manifold to obtain the new value of the extraction matrix

$$\hat{\mathbf{U}}^{[q]} = \mathbf{V}_L \mathbf{V}_R^T \quad \text{where} \quad [\mathbf{V}_L, \mathbf{\Lambda}, \mathbf{V}_R] = \text{svd}(\nabla_{\hat{\mathbf{U}}^{[q]}} \Phi). \tag{5.35}$$

Following a similar approach to the one presented in (Cruces, Cichocki and De Lathauwer, 2004), it can be shown at each of these iterations we guarantee the monotonous ascent in the contrast function.

Although this ICA method can easily recover a subset of the independent components from the observed signals, in our case, it is critical to select only those components which are related to the MI movements. We can easily favor their selection by initializing the unmixing matrix with the solution provided by the multiclass CSP algorithm. At the convergence of Eqn. 5.35 the projection matrix represents the spatial filters for four different MI movements.

The parameters of the ThinICA-CSP method can be tuned depending on the dataset, in our simulations, we have set them to: $p = 8$ (eight independent components had been extracted), $K = 3$ (three splits in the data has been considered), $\mathcal{T} = \{1, \ldots, 20\}$ (the set of delays), while the weighting terms have been set to $\gamma_4 = 0.025$ and $\gamma_2 = 1 - \gamma_4$. The main steps of the Sub-ABLD iteration are summarized in Algorithm 1.

Algorithm 1 ThinICA-CSP algorithm

function ThinICA-CSP$\{\mathbf{X}(t)\}$,$\mathbf{W}_{mCSP}$,p,K,$\mathcal{T}$

Compute the whitening transform matrix $\mathbf{T}$, $\mathbf{T} = (\mathbf{Cov}_x - \sigma_n^2\mathbf{I})^{\frac{1}{2}}X$.

Whiten the input data, $\mathbf{Z} = \mathbf{TX}$.

Initialize the semi-orthogonal matrix $\hat{\mathbf{U}}$ by computing the SVD of $\mathbf{W}_{mCSP}\mathbf{T}^{-1}$

Compute the estimated output, $\mathbf{Y}$ using Eqn. 5.30

Initialize the iteration counter: $i = 0$

repeat

    Compute the second order statistic

    Compute the fourth order statistic

    Compute the objective function using Eqn. 5.34

    Compute the gradient, $\nabla_{\hat{\mathbf{U}}^{[q]}}\Phi$

    The projection is done by thin SVD factorization of the gradient $\nabla_{\hat{\mathbf{U}}^{[q]}}\Phi$

    The matrix $\hat{\mathbf{U}}^{[q]}$ is obtained by using Eqn. 5.35

    Compute the estimated output

until convergence.

return $\mathbf{W}^T = \hat{\mathbf{U}}^T\mathbf{T}$

end function

5.3.4 Feature Extraction

In this study, 2 filters have been selected for each class which gives a total of $p = 8$ spatial filters for four classes.The training and testing signals were filtered using the obtained spatial filters. The log transformation of the variance ($var(.)$) of the spatially filtered signals gives the required features.

$$Feature_i = \log(\text{var}(\hat{\mathbf{u}}_i^T\mathbf{z})). \tag{5.36}$$

5.3.5 Classification

The extracted training features were used for training the classifiers. In order to perform a comparative analysis using the two classifiers, LDA and SVM were used for this study. In this study, the M-SVM^2 method from MSVM package (Lauer and Guermeur, 2011) is used for the discrimination of multiclass MI movements. The M-SVM^2 is the extended

version of 2-norm SVM for the multiclass approach. The detailed explanation can be found in (Guermeur and Monfrini, 2011). SVM is also applied in multiclass problem like LDA (Schlögl et al., 2005).

5.4 Results

LDA and SVM are commonly use classifiers in the discrimination of MI movements for BCI applications. Therefore, in order to perform a comparative study of these two classifiers using BCI competition IV dataset 2a, we have tested on mCSP algorithm and the proposed ThinICA-CSP algorithm. The performance observed for all the combinations are shown in Table 5.1 and Fig. 5.3. From the results, it is observed that LDA performs better than SVM for both mCSP and ThinICA-CSP algorithm. Therefore, LDA classifier is chosen for all the remaining studies.

Table 5.1 Comparative accuracy using LDA and SVM classifier with mCSP and ThinICA-CSP

Methods	Classifiers	Accuracy(%)
mCSP	LDA	62
	SVM	60
ThinICA-CSP	LDA	64
	SVM	63

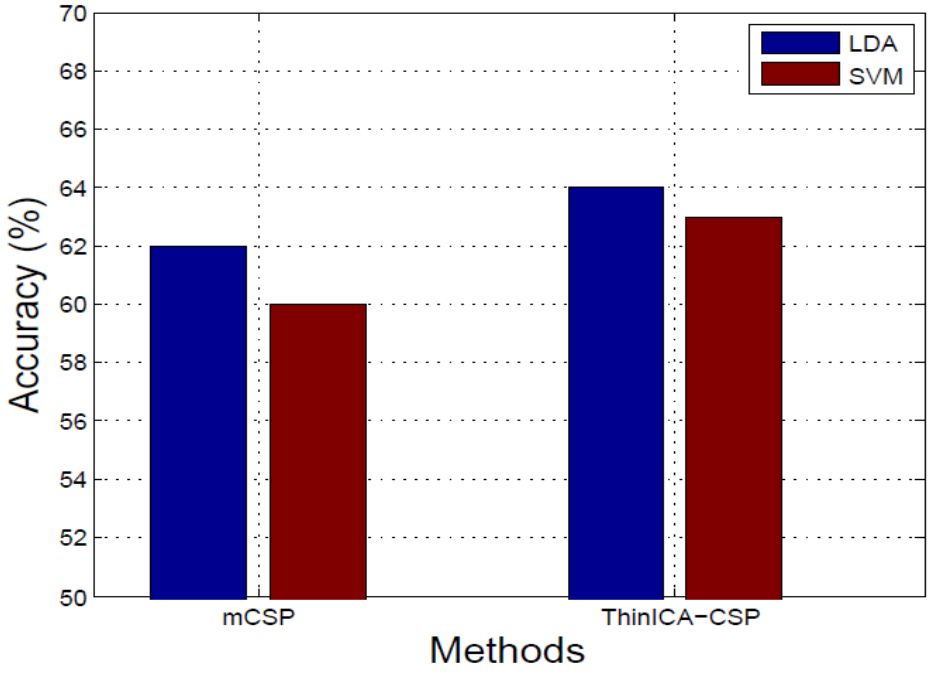


Fig. 5.3 Comparative analysis of LDA and SVM classifier using mCSP and ThinICA-CSP

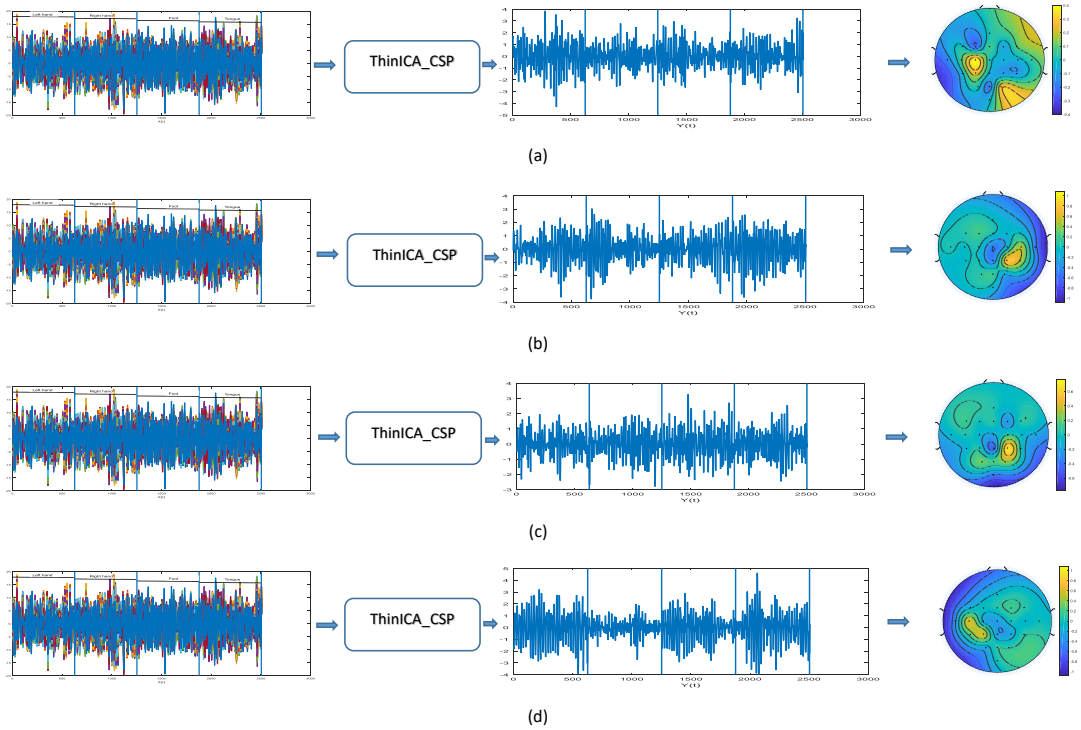


Fig. 5.4 Illustration of the four class (left hand, right hand, foot and tongue) MI movements processed with ThinICA-CSP algorithm, the corresponding filtered signals and spatial patterns for (a) left hand,(b) right hand, (c)foot and (d) tongue for subject A1.

Moreover, from the figure, it can also observe that ThinICA-CSP gives the highest performance of 64% than the multiclass CSP.

The processed EEG signals for subject A1 at each stage are shown in Fig. 5.4. The first row shows the input EEG signals for left hand, right hand, foot and tongue motor imagery signals which is processed using the proposed ThinICA-CSP algorithm. The third row represents the filtered signals and the fourth row shows the spatial patterns for four different MI movements.

5.5 Conclusions

ThinICA-CSP algorithm for discrimination of multiclass MI movements is proposed. The comparative study using LDA and SVM classifiers shows that LDA performs better than SVM for this scenario. Therefore, further studies are done using LDA. The experimented result shows that the proposed algorithm performs better than multiclass CSP algorithm. From the above observation, it can be concluded that utilization of second and higher order statistics and initialization of unmixing matrix with multiclass CSP filter matrix improves the classification performance of multiclass movements. Hence, this contribution provides additional evidence in favor of the use of ICA techniques for the robust classification of multiclass movements.

CHAPTER 6

Divergence maximization and its Relation with CSP

Over the last few years, the use of specialized metrics and divergences measures in the successful design of dimensionality reduction techniques has been progressively acquiring much recognition (Samek et al., 2014; Harandi et al., 2017; Huang et al., 2015; Horev et al., 2016). There are numerous real scenarios and applications for which the parameters of interest belong to non-flat manifolds, and where the Euclidean geometry results are unsuitable to evaluate the similarities. Indeed, this is the case in the comparison of probability density functions and also of their associated covariance matrices. The present contribution may be seen as a continuation of the work in (Cichocki et al., 2015), where we defined the Alpha-Beta Log-Det family of divergences between Symmetric and Positive Definite (SPD) matrices and studied its properties. The Alpha-Beta Log-Det family unifies under the same framework of many existing Log-Det divergences and connects them smoothly, through intermediate versions, with the help of two real hyperparameters: α and β. In (Quang, 2016) a recent extension of the Alpha-Beta Log-Determinant divergences was also proposed for the infinite-dimensional setting.

The evaluation of the Alpha-Beta Log-Det divergences depends on the generalized eigenvalues of the compared SPD matrices, and makes its optimization non-trivial. This motivates us to interpret the optimization of AB Log-Det divergence as CSP and derived the gradient for optimization.

This chapter is organized as follows: the related background is explained in section 6.1. Section 6.2 presents the fundamental model of the observations and notation. Section 6.3 reviews the CSP algorithm while section 6.4 discusses CSP via the divergence optimization. In section 6.5, we present the family of AB Log-Det divergences and provides new upper-bounds and conditions for the equivalence between this divergence optimization and the robust CSP solution. Section 6.6 explains how to obtain closed-form formulas for computing the gradient of the AB Log-Det divergence, which is useful for its optimization. The analysis of the robustness of the divergence in terms of its hyperparameters is the objective of section 6.7.

6.1 Background

Brain-Computer Interface has gained lots of interest in neuroscience and rehabilitation engineering. BCI (Dornhege, 2007; Wolpaw and Wolpaw, 2012) systems enable a person to operate external devices by using brain signals. The MI-based BCI systems are the preferable BCI systems among others. It uses the brain signals of the MI movements as control commands for external devices without using the peripheral nervous system. During the imagination process, an alteration in the rhythmic activity of the brain can be observed in the *mu* and *beta* rhythms at the corresponding area of the sensory-motor cortex. This phenomenon is known as ERS or ERD (Pfurtscheller and Da Silva, 1999). The MI-based BCI systems use these activities as control commands. Such a system can potentially serve as a communication aid for the people suffering from ALS, multiple sclerosis and completely locked-in.

As mentioned in the previous chapter, CSP algorithm is one of the most popular and efficient algorithms used for MI-based BCI applications (Fukunaga and KoonTz, 1970). It was first used to detect the abnormalities present in EEG signals (Koles, 1991) and later, was introduced in BCI applications (Ramoser et al., 2000). The main objective of the CSP is to obtain the spatial filters by maximizing the variance of one class, at the same time minimizing that of the other class variance. It has been reported that this algorithm provides excellent classification accuracy for MI-based BCI systems. Besides, being the most popular method, its performance is easily affected by the presence of artifacts and nonstationarities. Since the computation of the spatial filters mainly depends on the covariance matrix, the presence of artifacts such as blinking of the eyes, eye movements and improper placement of the electrodes contribute to the poor computation of the covariance matrix which leads to the poor classification performance.

The main contributions of this work are the following: The existing link between the CSP method and the symmetric KL divergence (see (Samek et al., 2014)), is extended to the case of the minimax optimization of the AB Log-Det divergences. In absence of regularization, their solutions are shown to be equivalent whenever these methods apply the same divergence-based criterion for choosing the spatial filters. Although, in general, this is not the case when the CSP method adopts the popular practical criterion of a priori fixing the number of spatial filters for each class, it is shown that the equivalence with the solution of the optimization of AB Log-Det divergences can be still preserved if a suitable scaling factor κ is used in one of the arguments of the divergence.

The details on how to perform the optimization of the AB Log-Det divergence are presented. The explicit expression of the gradient of this divergence with respect to the spatial filters is obtained. Expression which generalizes and extends the gradient of several more established well-known divergences, for instance, the gradient of the Alpha–Gamma divergence and the gradient of the Kullback–Leibler divergence between

SPD matrices.

6.2 Notation and Model of the Measurements

The following notations are adopted. Vectors are typically denoted by bold letters, the capital bold letters are reserved for the matrices, while the random variables appear in italic capital letters. The operators $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ round the value of their argument to the nearest lower and higher integers respectively. All the covariance matrices, which are denoted by $Cov(\cdot)$, are assumed to be positive definite and hence invertible.

Let us now describe the statistical model of the observations. As usual, the raw EEG observations are initially preprocessed by a bandpass filter that retains the activity in the bands of the mu and $beta$ rhythms and is later normalized for each trial so as to keep their total spatial power constant. One can define a statistical model of these "normalized" observations as $\boldsymbol{x}(t) = [x_1(t), \ldots, x_n(t)]^T$ conditioned on the true imagery movement, which here will be represented by a member of the class $c \in \{c_1, c_2\}$. In general, the EEG observations are noisy and high-dimensional, while the number of recorded trials is quite limited. Therefore, the learning of the discriminative features is quite sensitive to overfitting, a situation that would severely degrade the prediction accuracy over test samples. In this case, it is worth sacrificing the bias by choosing a simpler (less complex) model in which parameters can be estimated with a smaller variance. For this reason, we adopt usual convention (Wu et al., 2015) of considering the observations from each class as drawn from the independent and identically distributed (i.i.d.) Gaussian random vectors as represented as $\boldsymbol{X}|c$ of zero mean and with covariance matrix as $Cov(\boldsymbol{X}|c)$, which in turn is set equal to the sample covariance matrix of the class, i.e.,

$$Cov(\boldsymbol{X}|c) = Cov(\boldsymbol{x}|c) \quad \text{for} \quad c \in \{c_1, c_2\}. \tag{6.1}$$

The observations are then modeled by the mixture distribution

$$p(\boldsymbol{x}) = p(c_1)p(\boldsymbol{x}|c_1) + p(c_2)p(\boldsymbol{x}|c_2), \tag{6.2}$$

where $p(c)$ refers to the sample probabilities of each class in the training data. When $\bar{\boldsymbol{x}}$ denotes the sample mean of the observations, their sample covariance matrix is obtained by

$$Cov(\boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{x}(t) - \bar{\boldsymbol{x}}) (\boldsymbol{x}(t) - \bar{\boldsymbol{x}})^T = p(c_1)Cov(\boldsymbol{x}|c_1) + p(c_2)Cov(\boldsymbol{x}|c_2) \tag{6.3}$$

and its eigenvalue decomposition is

$$Cov(\boldsymbol{x}) = \mathbf{U}_1 \boldsymbol{\Delta} \mathbf{U}_1^T \tag{6.4}$$

where $\mathbf{\Delta}$ and $\mathbf{U}_1$, respectively denote the matrix of eigenvalues and eigenvectors of $Cov(\boldsymbol{x})$.

We define $\boldsymbol{w}_i = [w_{1i}, w_{2i}, ..., w_{pi}]^T$ as the vector with the coefficients of the i-th-esime spatial filter for $i = 1, \ldots, p$. The collection of p spatial filters forms the overall filter matrix $\mathbf{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2 ..... \boldsymbol{w}_p]$, which is used to reduce the dimensionality of the observations by projecting them onto the p-dimensional subspace spanned by the filter outputs

$$\boldsymbol{y} = \mathbf{W}^T \boldsymbol{x} \in \mathbb{R}^p, \tag{6.5}$$

where $p \ll n$. The model for the estimated conditional distribution $p(\boldsymbol{y}|c)$ is a multidimensional Gaussian of zero mean and covariance matrix $Cov(\boldsymbol{Y}|c) = \mathbf{W}^T Cov(\boldsymbol{x}|c)\mathbf{W}$, i.e., for each class

$$\boldsymbol{Y}|c \sim \mathcal{N}(\mathbf{0}, \mathbf{W}^T Cov(\boldsymbol{x}|c)\mathbf{W})\,. \tag{6.6}$$

6.3 The Common Spatial Patterns Algorithm

The development of the CSP algorithm as a technique for feature selection in classification problems can be traced back to the work of (Fukunaga and KoonTz, 1970), while later, (Koles, 1991; Ramoser et al., 2000) considered its practical application for the study of EEG signals. This technique exploits the event-related desynchronization during the limbs movement imagination process that alters the rhythmic activity in a class dependent area of the motor cortex. The objective of the algorithm is to obtain a set of most discriminative spatial filters, i.e., those that hierarchically maximize the output activity of one class, while at the same time; they minimize the activity of the other class. Since only the direction of the spatial filters (i.e., not the scale) are of interest, the technique starts with a linear transformation $\boldsymbol{y} = \mathbf{W}^T \boldsymbol{x}$ that whitens the sample covariance of the outputs

$$Cov(\boldsymbol{y}) = p(c_1)Cov(\boldsymbol{y}|c_1) + p(c_2)Cov(\boldsymbol{y}|c_2) = \mathbf{W}^T\, Cov(\boldsymbol{x})\, \mathbf{W} = \mathbf{I}_p\,. \tag{6.7}$$

With the help of the eigenvalue decomposition of $Cov(\boldsymbol{x})$, the general expression of the spatial filter matrix that preserves the whitening constraint can be found as

$$\mathbf{W}^T = \mathbf{\Omega}^T \mathbf{\Delta}^{-\frac{1}{2}} \mathbf{U}_1^T\,. \tag{6.8}$$

Note that $\mathbf{W} \in \mathbb{R}^{n \times p}$ is specified up to the ambiguity in the choice of the semi-orthogonal matrix $\mathbf{\Omega} \in \mathbb{R}^{n \times p}$ (i.e., $\mathbf{\Omega}^T \mathbf{\Omega} = \mathbf{I}_p$) which parameterizes the relevant degrees of freedom for finding the most discriminative directions. Then, the objective of the CSP criterion (Fukunaga and KoonTz, 1970) is implemented by first choosing one part of the spatial filters from the constrained maximization of the conditional covariances of the

outputs of the first class

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}} \boldsymbol{w}^T Cov(\boldsymbol{x}|c_1)\boldsymbol{w} \qquad i = 1, \ldots, k, \tag{6.9}$$

and later choosing the other part of the filters to hierarchically maximize the conditional covariances of the outputs of the second class

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}} \boldsymbol{w}^T Cov(\boldsymbol{x}|c_2)\boldsymbol{w} \qquad i = k+1, \ldots, p, \tag{6.10}$$

where, in both cases, the maximization with respect to the spatial filters takes place under the whitening or ($Cov(\boldsymbol{x})$-orthonormality) constraints

$$\boldsymbol{w}_i^T Cov(\boldsymbol{x})\boldsymbol{w}_j = \delta_{ij} \qquad \forall j \leq i\,. \tag{6.11}$$

The number of spatial filters k that hierarchically maximize Eqn. 6.9 can be determined by a chosen filter selection policy. For simplicity, in most cases it is usually set k close to $\frac{p}{2}$ with the aim to balance the number of spatial filters devoted to each of the classes.

The maximization in Eqn. 6.9 can be alternatively posed as the constrained optimization of the quotient

$$\boldsymbol{w}_i = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^T Cov(\boldsymbol{x}|c)\boldsymbol{w}}{\boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w}} \quad \text{subject to} \quad \boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w} = \delta_{ij} \quad \forall j \leq i \tag{6.12}$$

which, in terms of the transformed and normalized spatial vectors

$$\boldsymbol{r}_i = \frac{(Cov(\boldsymbol{x}))^{\frac{1}{2}}\boldsymbol{w}_i}{||(Cov(\boldsymbol{x}))^{\frac{1}{2}}\boldsymbol{w}_i||_2}, \tag{6.13}$$

is rewritten as a quadratic optimization under orthogonality constraints

$$\begin{aligned} \boldsymbol{w}_i = (Cov(\boldsymbol{x}))^{-\frac{1}{2}} \times \arg\max_{\boldsymbol{r}} \left\{ \boldsymbol{r}^T \, (Cov(\boldsymbol{x}))^{-\frac{1}{2}} Cov(\boldsymbol{x}|c)(Cov(\boldsymbol{x}))^{-\frac{1}{2}} \, \boldsymbol{r} \right\} \\ \text{s.t.} \quad \boldsymbol{r}^T \boldsymbol{r}_j = \delta_{ij} \quad \forall j \leq i\,. \end{aligned} \tag{6.14}$$

At this point, the straightforward application of the Courant–Fisher–Weyl minimax principle ((Bhatia, 1997), p. 58) yields the variational description of the desired spatial

filters as the minimax solution of the Rayleigh quotients for each class

$$\boldsymbol{w}_i = (Cov(\boldsymbol{x}))^{-\frac{1}{2}} \times \arg \min_{\dim\{\mathcal{R}\}=n-i+1} \max_{\substack{\boldsymbol{r}\in\mathcal{R} \\ \|\boldsymbol{r}\|=1}} \left\{ \boldsymbol{r}^T \left(Cov(\boldsymbol{x})\right)^{-\frac{1}{2}} Cov(\boldsymbol{x}|c)(Cov(\boldsymbol{x}))^{-\frac{1}{2}} \boldsymbol{r} \right\} \tag{6.15}$$

$$= \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} \frac{\boldsymbol{w}^T Cov(\boldsymbol{x}|c)\boldsymbol{w}}{\boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w}} \tag{6.16}$$

$$= \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} \frac{Cov(y_i|c)}{Cov(y_i)}. \tag{6.17}$$

By the same principle, the generalized eigenvectors $\boldsymbol{v}_i^{(c)}$ of the matrix pencil $(p(c)\ Cov(\boldsymbol{y}|c),\ Cov(\boldsymbol{y}))$, are the minimax solutions of the Rayleigh quotient, while the values that takes the criterion at these solutions are the generalized eigenvalues

$$\lambda_i^{(c)} = p(c)\frac{\boldsymbol{v}_i^{(c)T} Cov(\boldsymbol{x}|c)\boldsymbol{v}_i^{(c)}}{\boldsymbol{v}_i^{(c)T} Cov(\boldsymbol{x})\boldsymbol{v}_i^{(c)}} = p(c) \min_{\dim\{\mathcal{W}\}=i} \max_{\boldsymbol{w}\in\mathcal{W}} \frac{\boldsymbol{w}^T Cov(\boldsymbol{x}|c)\boldsymbol{w}}{\boldsymbol{w}^T Cov(\boldsymbol{x})\boldsymbol{w}}, \tag{6.18}$$

which are sorted according to the descent in their magnitude, $\lambda_1^{(c)} \geq \lambda_2^{(c)} \geq \ldots \geq \lambda_n^{(c)}$.

The generalized eigenvectors of the two quotients (one for each class) coincide, except for their ordering which are reversed (Fukunaga and KoonTz, 1970), i.e., $\boldsymbol{v}_i^{(c_1)} = \boldsymbol{v}_{n-i+1}^{(c_2)}$, while the weighted sum of generalized eigenvalues is bounded by

$$\lambda_i^{(c_1)} + \lambda_{n-i+1}^{(c_2)} = \frac{\boldsymbol{v}_i^{(c_1)T} \left(p(c_1)Cov(\boldsymbol{x}|c_1) + p(c_2)Cov(\boldsymbol{x}|c_2)\right) \boldsymbol{v}_i^{(c_1)}}{\boldsymbol{v}_i^{(c_1)T} Cov(\boldsymbol{x})\boldsymbol{v}_i^{(c_1)}} = 1. \tag{6.19}$$

Therefore, a direction of maximum variance for one class will simultaneously minimize the variance of the other class, and vice versa. Hence, the standard CSP solution is obtained when the spatial filters match with the principal and minor eigenvectors of the generalized eigendecomposition problem (Fukunaga and KoonTz, 1970; Koles, 1991; Ramoser et al., 2000)

$$Cov(\boldsymbol{x}|c_1)\boldsymbol{v}_i^{(c_1)} = \lambda_i^{(c_1)} Cov(\boldsymbol{x})\boldsymbol{v}_i^{(c_1)} \quad i = 1, \ldots, n. \tag{6.20}$$

After sorting the eigenvalues according to its magnitude, CSP explicitly selects k spatial filters $\boldsymbol{v}_i^{(c_1)}$ from the principal eigenvectors and $p-k$ spatial filters from the minor eigenvectors, to form the spatial filter matrix

$$\mathbf{W}_{CSP} \equiv [\boldsymbol{w}_1, \boldsymbol{w}_2 ..... \boldsymbol{w}_p] = [\boldsymbol{v}_1^{(c_1)}, \ldots, \boldsymbol{v}_k^{(c_1)}, \boldsymbol{v}_{n-(p-k)+1}^{(c_1)}, \ldots, \boldsymbol{v}_n^{(c_1)}]. \tag{6.21}$$

6.4 The Divergence Optimization Interpretation of CSP

Under the appropriate selection policy for the number of spatial filters for each class, the solution obtained by the CSP algorithm admits an interpretation in terms of the optimization divergence measures (here denoted by $Div(\cdot\|\cdot)$) between the Gaussian pdfs outputs for each class

$$\boldsymbol{w}_i = \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} Div(p(y_i|c_1)\|p(y_i|c_2))\,, \tag{6.22}$$

except for a probable permutation in the ordering of some of the spatial filters.

The problem can be formulated using the following optimization problem

$$\boldsymbol{w}_i = \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} D(Cov(y_i|c_1)\|Cov(y_i|c_2)) \tag{6.23}$$

where $D(\cdot\|\cdot)$ refers to a divergence between the covariances of the conditional densities of the outputs. As a consequence of the assumption of zero mean Gaussian densities, the covariances are the only necessary statistics that summarize all the relevant information of the conditional data.

In particular, the solution of the CSP algorithm was linked in (Fukunaga and KoonTz, 1970; Samek et al., 2014; Wang, 2012; Samek, Blythe, Müller and Kawanabe, 2013) with the optimization of the symmetric Kullback–Leibler divergence (sKL)

$$Div_{sKL}(p(y_i|c_1)\|p(y_i|c_2)) = \int p(y_i|c_1)\log\frac{p(y_i|c_1)}{p(y_i|c_2)}dy_i + \int p(y_i|c_2)\log\frac{p(y_i|c_2)}{p(y_i|c_1)}dy_i, \tag{6.24}$$

$$= \int (p(y_i|c_1) - p(y_i|c_2))\log\frac{p(y_i|c_1)}{p(y_i|c_2)}dy_i. \tag{6.25}$$

This divergence measures can be simplified to the symmetric Kullback–Leibler (sKL) divergence between the class conditional covariances

$$Div_{sKL}(p(y_i|c_1)\|p(y_i|c_2)) = \frac{1}{2}\frac{Cov(y_i|c_1)}{Cov(y_i|c_2)} + \frac{1}{2}\frac{Cov(y_i|c_2)}{Cov(y_i|c_1)} - 1 \tag{6.26}$$

$$\equiv D_{sKL}(Cov(y_i|c_1)\|Cov(y_i|c_2))\,. \tag{6.27}$$

In this work, an extension of the existing KL to the criterion of the AB Log-Det divergence ($D_{AB}^{(\alpha,\beta)}(.\|.)$) between the class-conditional covariances defined as (Cichocki

et al., 2015) is proposed

$$D_{AB}^{(\alpha,\beta)}(Cov(y_i|c_1)\|Cov(y_i|c_2)) = \frac{1}{\alpha\beta}\log\left|\frac{\alpha\left(\frac{Cov(y_i|c_1)}{Cov(y_i|c_2)}\right)^{\beta}+\beta\left(\frac{Cov(y_i|c_2)}{Cov(y_i|c_1)}\right)^{\alpha}}{\alpha+\beta}\right|_{+} \quad \text{for } \alpha\neq 0,\ \ \beta\neq 0,\ \ \alpha+\beta\neq 0, \tag{6.28}$$

where

$$|x|_{+}=\begin{cases} x & x\geq 0,\\ 0, & x<0,\end{cases} \tag{6.29}$$

denotes the non-negative truncation operator. When the arguments covariances are scalars and $\alpha, \beta > 0$, the AB Log-Det divergence can also be rewritten as the logarithmic ratio between the weighted arithmetic mean of the scaled covariances ($Cov^{\alpha+\beta}(y_i|c_1)$, $Cov^{\alpha+\beta}(y_i|c_2)$) and their weighted geometric mean, i.e.,

$$D_{AB}^{(\alpha,\beta)}(Cov(y_i|c_1)\|Cov(y_i|c_2)) = \tfrac{1}{\alpha\beta}\log\frac{\left(\frac{\alpha}{\alpha+\beta}Cov^{\alpha+\beta}(y_i|c_2)+\frac{\beta}{\alpha+\beta}Cov^{\alpha+\beta}(y_i|c_1)\right)}{(Cov^{\alpha+\beta}(y_i|c_2))^{\frac{\alpha}{\alpha+\beta}}(Cov^{\alpha+\beta}(y_i|c_1))^{\frac{\beta}{\alpha+\beta}}}. \tag{6.30}$$

Additionally if $\alpha+\beta=1$, the AB Log-det divergence between covariances is proportional to the Alpha–Gamma divergence $D_{AB}^{(\alpha,\beta)}(.\|.)$ (Cichocki, 2010) between the conditional densities

$$\begin{aligned} D_{AB}^{(\alpha,\beta)}(Cov(y_i|c_1)\|Cov(y_i|c_2)) &\equiv 2\ Div_{AG}^{(\beta,\alpha)}\left(p(y_i|c_1)\|p(y_i|c_2)\right)\\ &= \frac{2}{\alpha\beta}\log\frac{\left(\int_{\Omega}p(y_i|c_1)\ dy_i\right)^{\beta}\left(\int p(y_i|c_2)\ dy_i\right)^{\alpha}}{\int p^{\beta}(y_i|c_1)\ p^{\alpha}(y_i|c_2)\ dy_i} \end{aligned} \tag{6.31}$$

$$\text{for}\quad \alpha>0,\ \beta>0,\ \alpha+\beta=1.$$

In Section 6.5.3, it is proven that, under certain conditions, the simple optimization of an AB Log-Det divergence also leads to the solution of the CSP algorithm. Although, the potential of these divergences does not rely on their plain optimization but instead rely on their optimization in the presence of some regularization terms that help to specify the desired solutions.

Recently, several divergence criteria have been proposed for the extraction of the spatial dimensions with maximum discriminative power. Among these, the multiclass approach based on the maximization of the harmonic mean of Kullback–Leibler divergences (Wang, 2012) and the regularization framework based on the beta divergences

(Samek et al., 2014; Samek, Blythe, Müller and Kawanabe, 2013) are the most noteworthy methods. Another approach based on Bhattacharyya distance and Gamma divergence has also been proposed for classification of motor imagery movements (Brandl et al., 2015). Our work is motivated by the success of these methods in improving the classification accuracy and the robustness against the outliers. The distinctive property of the AB Log-Det divergence is that it smoothly connects (through its hyperparameters) a quite broad family of Log-Det divergences for SPD matrices, covering several relevant classical cases like: the KL divergence, the dual KL divergence, the Beta Log-Det family, the Alpha Log-Det family, the Power Log-Det family, as well as the Affine Invariant Riemannian divergence.

6.5 The Definition of the AB Log-Det Divergence

Henceforth, we will work on the multidimensional observation vectors $\boldsymbol{x} = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$. In order to simplify the notation, the covariance matrices of the two classes are renamed as follows

$$\mathbf{P} \equiv Cov(\boldsymbol{x}|c_1), \mathbf{Q} \equiv Cov(\boldsymbol{x}|c_2). \tag{6.32}$$

The AB Log-Det divergence is a directed divergence that evaluates the dissimilarity between two multidimensional covariance matrices. It was defined in (Cichocki et al., 2015) as

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{\alpha\beta}\log\left|\frac{\alpha(\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}})^{\beta} + \beta(\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}})^{-\alpha}}{\alpha+\beta}\right|_{+} \tag{6.33}$$

$$\text{for } \alpha \neq 0, \quad \beta \neq 0, \quad \alpha+\beta \neq 0,$$

while, for the singular cases, its definition is given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \begin{cases} \frac{1}{\alpha^2}\left[\mathrm{tr}\left((\mathbf{Q}^{\frac{1}{2}}\mathbf{P}^{-1}\mathbf{Q}^{\frac{1}{2}})^{\alpha} - \mathbf{I}\right) - \alpha\log|\mathbf{Q}^{\frac{1}{2}}\mathbf{P}^{-1}\mathbf{Q}^{\frac{1}{2}}|\right] & \text{for } \alpha \neq 0,\ \beta = 0, \\ \frac{1}{\beta^2}\left[\mathrm{tr}\left((\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}})^{\beta} - \mathbf{I}\right) - \beta\log|\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}}|\right] & \text{for } \alpha = 0,\ \beta \neq 0, \\ \frac{1}{\alpha^2}\log\left|(\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}})^{\alpha}(\mathbf{I} + \log(\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}})^{-\alpha})\right|_{+} & \text{for } \alpha = -\beta, \\ \frac{1}{2}||\log(\mathbf{Q}^{\frac{1}{2}}\mathbf{P}^{-1}\mathbf{Q}^{\frac{1}{2}})||_F^2 & \text{for } \alpha,\ \beta = 0. \end{cases} \tag{6.34}$$

The divergence depends only on the eigenvalues $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ of the SPD matrix $\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2}$, which also coincide with the eigenvalues of the matrix $\mathbf{Q}^{-1}\mathbf{P}$, although

their eigenspaces differ. Given the eigenvalue decomposition

$$\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}} = \mathbf{V}_1\mathbf{\Lambda}\mathbf{V}_1^T, \tag{6.35}$$

where $\mathbf{V}_1$ is the orthogonal matrix of eigenvectors, and $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is the diagonal matrix with positive eigenvalues $\lambda_i > 0,\ i = 1, 2, \ldots, n$. One of the properties of the AB Log-Det divergence is that it is invariant under a common change of basis on its matrix arguments, i.e., an invertible congruence transformation. Since, with the help of this specific transformation, we have

$$\mathbf{P} \to (\mathbf{V}_1^T\mathbf{Q}^{-\frac{1}{2}})\mathbf{P}(\mathbf{V}_1^T\mathbf{Q}^{-\frac{1}{2}})^T = \mathbf{\Lambda}, \mathbf{Q} \to (\mathbf{V}_1^T\mathbf{Q}^{-\frac{1}{2}})\mathbf{Q}(\mathbf{V}_1^T\mathbf{Q}^{-\frac{1}{2}})^T = \mathbf{I}, \tag{6.36}$$

it can be inferred that the divergence is separable (over the generalized eigenvalues of the matrix pencil $(\mathbf{P}, \mathbf{Q})$) in a sum of marginal divergences that measure how far are each of the generalized eigenvalues from the unity, i.e.,

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\ \|\mathbf{Q}) = D_{AB}^{(\alpha,\beta)}\left(\mathbf{\Lambda}\|\ \mathbf{I}_n\right) = \sum_{i=1}^{n} D_{AB}^{(\alpha,\beta)}(\lambda_i\|\ 1). \tag{6.37}$$

Hence,

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{\alpha\beta}\sum_{i=1}^{n}\log\left|\frac{\alpha\lambda_i^{\beta} + \beta\lambda_i^{-\alpha}}{\alpha+\beta}\right|_+, \quad \alpha,\ \beta,\ \alpha+\beta \neq 0. \tag{6.38}$$

Similarly, for the singular cases, the divergence is

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \begin{cases} \dfrac{1}{\alpha^2}\left[\displaystyle\sum_{i=1}^{n}\left(\lambda_i^{-\alpha} - \log(\lambda_i^{-\alpha})\right) - n\right] & \text{for} \quad \alpha \neq 0,\ \beta = 0 \\ \dfrac{1}{\beta^2}\left[\displaystyle\sum_{i=1}^{n}\left(\lambda_i^{\beta} - \log(\lambda_i^{\beta})\right) - n\right] & \text{for} \quad \alpha = 0,\ \beta \neq 0 \\ \dfrac{1}{\alpha^2}\left[\displaystyle\sum_{i=1}^{n}\log\left|\frac{\lambda_i^{\alpha}}{1+\log\lambda_i^{\alpha}}\right|_+\right] & \text{for} \quad \alpha = -\beta \neq 0 \\ \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\log^2(\lambda_i) & \text{for} \quad \alpha,\ \beta = 0. \end{cases} \tag{6.39}$$

This divergence compares two symmetric positive definite matrices and returns its dissimilarity, i.e., a positive value when they are non-coincident and $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = 0$ iff $\mathbf{P} = \mathbf{Q}$. As it can be observed in Fig. 6.1 the AB Log-Det divergence generalizes

several existing log-det matrix divergences, like: the Steins loss, the S-divergence, the Alpha and Beta log-det families of divergences and the geodesic distance between covariance matrices (the squared Riemannian metric), among others (see Table 1 in (Cichocki et al., 2015) for a comprehensive list).

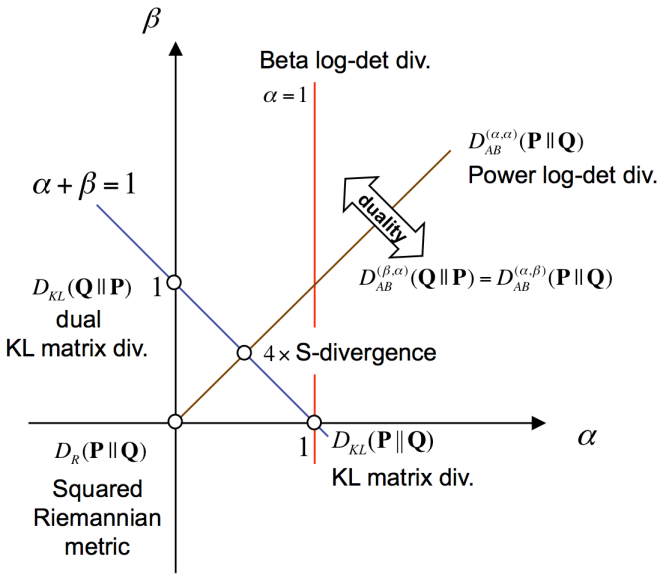


Fig. 6.1 This illustration shows the AB Log-Det divergence $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ positioned in a plane as a function of their real pair of hyperparameters (α, β). It is clear from the figure, that the parameterization smoothly connects several relevant positive definite matrix divergences, like: the squared Riemannian metric ($\alpha = 0, \beta = 0$), the KL matrix divergence or Stein's loss ($\alpha = 1, \beta = 0$), the dual KL matrix divergence ($\alpha = 0, \beta = 1$), and the S-divergence ($\alpha = \frac{1}{2}, \beta = \frac{1}{2}$) among others.

6.5.1 A Tight Upper-Bound for the AB Log-Det Divergences

The divergence $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ depends on the generalized eigenvalues $\lambda_1, \cdots, \lambda_n$ of the matrix pencil $(\mathbf{P}, \mathbf{Q})$ which, for convenience, are assumed to have a simple spectrum (the eigenvalues are unique or non-coincident) and can be sorted in descending order

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0. \tag{6.40}$$

In practice, the assumption is plausible because the real symmetric matrices with unique eigenvalues are known to form an open dense set in the space of all the real symmetric matrices (Tao, 2012).

Although the space of the observations is high-dimensional, most of the discriminative information between the two conditions is confined into a low-dimensional subspace. Thus, the spatial filter matrix $\mathbf{W} \in \mathbb{R}^{n\times p}$ is used to reduce the dimensionality of the samples from n to p with the linear compression transformation $\boldsymbol{y} = \mathbf{W}^T\boldsymbol{x} \in \mathbb{R}^p$. It is shown in (Cichocki et al., 2015) that, after applying this compression to the arguments

of the divergence, the resulting output covariance matrices $\mathbf{W}^T\mathbf{P}\mathbf{W}$ and $\mathbf{W}^T\mathbf{Q}\mathbf{W}$ are more similar than in the original space, as shown in the below equation

$$D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\,\|\mathbf{W}^T\mathbf{Q}\mathbf{W}) = \sum_{i=1}^{p} D_{AB}^{(\alpha,\beta)}(\mu_i\|1) \leq \sum_{i=1}^{n} D_{AB}^{(\alpha,\beta)}(\lambda_i\|1) = D_{AB}^{(\alpha,\beta)}(\mathbf{P}\,\|\mathbf{Q}), \tag{6.41}$$

where $\mu_1 \geq \cdots \geq \mu_p > 0$ are the generalized eigenvalues of the matrix pencil $(\mathbf{W}^T\mathbf{P}\mathbf{W}, \mathbf{W}^T\mathbf{Q}\mathbf{W})$. However, this upper bound is loose for the case of interest (dimensionality reduction), i.e., when $p < n$. In Appendix A, the possible way to tighten the previous upper-bound with the following new proposal is shown

$$D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\,\|\mathbf{W}^T\mathbf{Q}\mathbf{W}) \leq \sum_{i=1}^{p} D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_i}\|\,1), \tag{6.42}$$

where π defines the permutation of the indices $1,\ldots,n$ that sorts the divergence of the eigenvalues from the unity in descending order

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_1}\|\,1) \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_2}\|\,1) \geq \cdots \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_n}\|\,1). \tag{6.43}$$

Moreover, the equality with the upper-bound is only obtained for those extraction matrices $\mathbf{W}$ that lie within the span of the p generalized eigenvectors of the matrix pencil $(\mathbf{P},\mathbf{Q})$ which are associated with the eigenvalues $\lambda_{\pi_1},\ldots,\lambda_{\pi_p}$ that maximize the divergence from unity in Eqn. 6.43.

6.5.2 Relationship between the Generalized Eigenvalues and Eigenvectors of the Matrix Pencils $(\mathbf{P},\mathbf{Q})$ and $(p(c_1)\mathbf{P}, Cov(\mathbf{x}))$

We have seen in the previous section that the tight upper-bound of the divergence is attained by a subset of the generalized eigenvectors of the matrix pencil $(\mathbf{P},\mathbf{Q})$, whereas, the CSP solution in Eqn. 6.21 depends on a subset of the generalized eigenvectors of another matrix pencil $(p(c_1)\mathbf{P}, Cov(\mathbf{x}))$. In this section, the close relationship between both eigendecompositions is addressed. For this purpose, $\mathbf{\Lambda}$ is denoted as the matrix of eigenvalues of $\mathbf{Q}^{-1}\mathbf{P}$ and $\mathbf{\Lambda}^{(c_1)}$ as the matrix of eigenvalues of $(Cov(\boldsymbol{x}))^{-1}\,p(c_1)\mathbf{P}$. Then, we write

$$(p(c_2)\mathbf{Q})^{-1}\,(p(c_1)\mathbf{P}) = [(Cov(\boldsymbol{x}))^{-1}(p(c_2)\mathbf{Q})]^{-1}\,[(Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})], \tag{6.44}$$

and use the decomposition of $Cov(\boldsymbol{x})$ in Eqn. 6.3 to substitute $p(c_2)\mathbf{Q} = Cov(\boldsymbol{x}) - p(c_1)\mathbf{P}$ in the previous equation. In this way, we obtain

$$(p(c_2)\mathbf{Q})^{-1}(p(c_1)\mathbf{P}) = [\mathbf{I}_n - (Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})]^{-1}\,[(Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})]. \tag{6.45}$$

The matrix of eigenvectors $\mathbf{V}$ of $\mathbf{Q}^{-1}\mathbf{P}$ diagonalizes both sides of the previous equation

$$\begin{aligned}
\mathbf{\Lambda}\frac{p(c_1)}{p(c_2)} &= \mathbf{V}^{-1}[\frac{p(c_1)}{p(c_2)}\mathbf{Q}^{-1}\mathbf{P}]\mathbf{V} && (6.46)\\
&= (\mathbf{V}^{-1}[\mathbf{I}_n - (Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})]^{-1}\mathbf{V})\ (\mathbf{V}^{-1}[(Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})]\mathbf{V}) && (6.47)\\
&= [\mathbf{I}_n - \mathbf{V}^{-1}(Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P}\mathbf{V})]^{-1}\ (\mathbf{V}^{-1}[(Cov(\boldsymbol{x}))^{-1}(p(c_1)\mathbf{P})]\mathbf{V}) && (6.48)\\
&= (\mathbf{I}_n - \mathbf{\Lambda}^{(c_1)})^{-1}\mathbf{\Lambda}^{(c_1)}. && (6.49)
\end{aligned}$$

Hence, the explicit relationship between the two sets of eigenvalues is obtained

$$\lambda_i\frac{p(c_1)}{p(c_2)} = \frac{\lambda_i^{(c_1)}}{1-\lambda_i^{(c_1)}} \equiv g(\lambda_i^{(c_1)}), \qquad i = 1,\ldots,n, \tag{6.50}$$

where $g(\lambda_i^{(c_1)})$, as can be seen in Fig. 6.2, is a strictly monotonous ascending function over the domain of $\lambda_i^{(c_1)} \in (0,1)$. Moreover, the Equations (6.46)–(6.49) imply that the matrix $\mathbf{V}$ of generalized eigenvectors of the matrix pencil $(\mathbf{P},\mathbf{Q})$ exactly coincides with the matrix $\mathbf{V}^{(c_1)}$ of generalized eigenvectors of the other matrix pencil $(p(c_1)\mathbf{P}, Cov(\mathbf{x}))$.

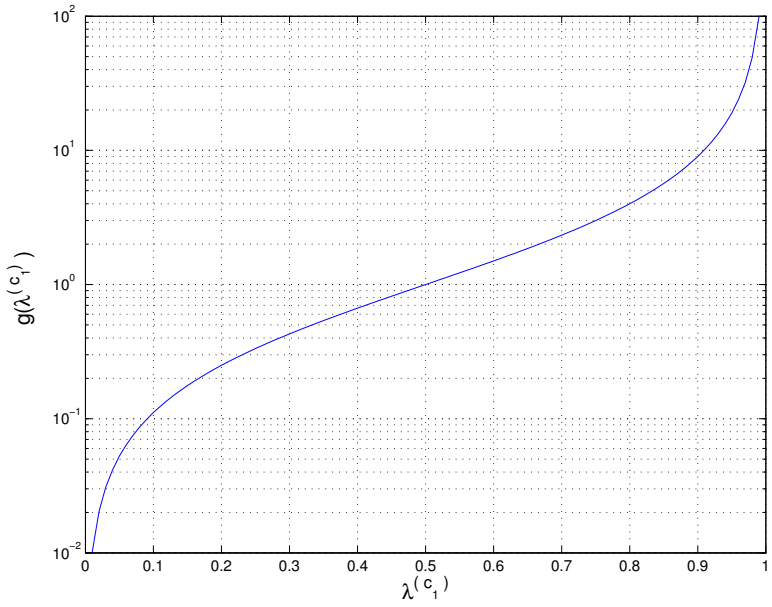


Fig. 6.2 Illustration of the strictly monotonous ascending transformation $g(\cdot)$ that, through Eqn. 6.50, maps eigenvalues of the matrix pencil $(p(c_1)\mathbf{P}, Cov(\mathbf{x}))$ into the eigenvalues of the matrix pencil $(\mathbf{P},\mathbf{Q})$, in a case where the sample probabilities of the classes are uniform $p(c_1) = p(c_2)$. Note that the eigenvalues of the first pencil are bounded in the interval $(0,1)$, while the domain of the eigenvalues of the second pencil is $(0,\infty)$.

6.5.3 Linking the Optimization of the Divergence and the CSP Solution

There is a link between the solutions of the CSP method and the solutions obtained with the optimization of the symmetric KL divergence between the class conditional

covariances, which was studied in previous works (Fukunaga and KoonTz, 1970; Samek et al., 2014; Wang, 2012). This subsection shows that under the appropriate filter selection criteria the link also extends to the optimization of other divergences, like the AB Log-Det family of divergences.

We have previously assumed that generalized eigenvalues are ordered and can be regarded as non-equal. Therefore, they can be clustered in the following three sets of principal, inner and minor eigenvalues of the matrix pencil $(\mathbf{P}, \mathbf{Q})$:

$$\underbrace{\lambda_1 > \cdots > \lambda_k}_{k \text{ principal eigenvalues}} > \underbrace{\lambda_{k+1} > \cdots > \lambda_{n-(p-k)}}_{\text{inner eigenvalues}} > \underbrace{\lambda_{n-(p-k)+1} > \cdots > \lambda_n}_{(p-k) \text{ minor eigenvalues}}. \tag{6.51}$$

The following sequence of optimizations induces an alternative ordering of the generalized eigenvalues

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_i} \| \, 1) = \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T \mathbf{P} \boldsymbol{w}_i \parallel \boldsymbol{w}_i^T \mathbf{Q} \boldsymbol{w}_i), \qquad i = 1, \ldots, n, \tag{6.52}$$

according to a permutation π that sorts their marginal divergences from 1 in descending order

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_1} \| \, 1) \geq \cdots \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_p} \| \, 1) \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_{p+1}} \| \, 1) \cdots \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_n} \| \, 1). \tag{6.53}$$

For building the matrix of spatial filters $\mathbf{W}_{Div} \equiv [\boldsymbol{w}_1, \boldsymbol{w}_2 ..... \boldsymbol{w}_p]$, one possible selection policy is to retain only the p most discriminative spatial filters for the considered divergence optimization problem, i.e., those that solve Eqn. 6.52 for $i = 1, \ldots, p$. The filters consist in p eigenvectors ($\boldsymbol{v}_{\pi_i}$ with $i = 1, \ldots, p$) of the matrix pencil $(\mathbf{P}, \mathbf{Q})$ that are arranged according to the permutation π. From the one-to-one relationship that exists between the generalized eigenvalues and eigenvectors of the matrix pencils $(\mathbf{P}, \mathbf{Q})$ and $(p(c_1)\mathbf{P}, Cov(\mathbf{x}))$ (see the previous subsection) the solution takes the following form

$$\mathbf{W}_{Div} = [\boldsymbol{v}_{\pi_1}, \ldots, \boldsymbol{v}_{\pi_p}] = [\boldsymbol{v}_{\pi_1}^{(c_1)}, \ldots, \boldsymbol{v}_{\pi_p}^{(c_1)}]. \tag{6.54}$$

This result tells us that the optimization of different divergences (in absence of other regularizing terms) only differs in the selection criteria for the spatial filters, which eventually determine the chosen subindices $\pi_1, \ldots, \pi_p$.

Now, the question of whether these spatial filters that solve the sequence of minimax divergence optimization problems

$$\min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T \mathbf{P} \boldsymbol{w}_i \parallel \boldsymbol{w}_i^T \mathbf{Q} \boldsymbol{w}_i), \qquad i = 1, \ldots, p, \tag{6.55}$$

essentially coincide (up to a possible permutation in the order of the spatial filters) with

the spatial filters of the CSP solution in Eqn. 6.56

$$\mathbf{W}_{CSP} = [\underbrace{\boldsymbol{v}_1^{(c_1)}, \ldots, \boldsymbol{v}_k^{(c_1)}}_{k \text{ principal eigenvectors}}, \underbrace{\boldsymbol{v}_{n-(p-k)+1}^{(c_1)}, \ldots, \boldsymbol{v}_n^{(c_1)}}_{p-k \text{ minor eigenvectors}}], \tag{6.56}$$

has a simple answer. The straightforward comparison between Eqn. 6.54 and Eqn. 6.56 reveals that both solutions should essentially coincide when the subindices $\pi_1, \ldots, \pi_p$ are a permutation of the integers $1, \ldots, k, n-(p-k)+1, \ldots, n$. Thus, the link between both techniques happens whenever CSP method adopts the filter selection policy of the divergence criterion in Eqn. 6.53.

However, many of the CSP implementations find satisfactory to choose the number of spatial filters for each class a priori, respectively as k and $p-k$ (we will refer to this case as the original CSP filter selection policy), where k is close to $p/2$ in order to approximately balance the number of spatial filters for each class (Blankertz et al., 2007; Ramoser et al., 2000).

In general, the use of a divergence based selection policy does not ensure a balanced representation of the spatial filters for each of the classes. For instance, consider the synthetic but illustrative situation for $n = 100$, where we wish to select $p = 8$ spatial filters. If the generalized eigenvalues of the matrix pencil $(\mathbf{P}, \mathbf{Q})$ are shifted towards to zero, for instance, equal to $\{10, 0.99, 0.98, \ldots, 0.03, 0.02, 0.01\}$. In most cases, the solution $\mathbf{W}_{Div}$ will select as its columns: only $k = 1$ principal eigenvectors and $p-k = 7$ minor eigenvectors, an unbalanced choice.

In view of this potential limitation, an interesting question is whether it would be possible to modify the AB Log-Det divergence criterion so as to enforce that its solution essentially coincides with the one obtained by the CSP method with its original filter selection policy. We will show in the following that this requires only a suitable scaling $\kappa \in \mathbb{R}^+$ in one of the arguments of the divergence. Without loss of generality, we assume scaling in the second argument of the divergence. As it is shown in the Appendix B, there is a permutation π' of the indices of the spatial filters $1, \ldots, p$ that links the CSP solution in Eqn. 6.21 with the optimization of the divergence

$$\boldsymbol{w}_{\pi_i'} = \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w} \in \mathcal{W}} D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T \mathbf{P} \boldsymbol{w}_i \parallel \kappa\, \boldsymbol{w}_i^T \mathbf{Q} \boldsymbol{w}_i), \qquad i = 1, \ldots, p, \tag{6.57}$$

for any given

$$\kappa \in (\kappa_{\inf}\ ,\ \kappa_{\sup}) \tag{6.58}$$

with

$$\kappa_{\inf} \equiv \mathcal{K}(\lambda_{k+1}\ ,\ \lambda_{n-(p-k)+1}) \kappa_{\sup} \equiv \mathcal{K}(\lambda_k\ ,\ \lambda_{n-(p-k)}) \tag{6.59}$$

where the function

$$\mathcal{K}(a,b) = \begin{cases} \left(\dfrac{(a^{\beta}-b^{\beta})/\beta}{(a^{-\alpha}-b^{-\alpha})/(-\alpha)}\right)^{\frac{1}{\alpha+\beta}} & \text{for } \alpha,\beta,\alpha+\beta \neq 0 \\ \left(\dfrac{\log(a/b))}{(a^{-\alpha}-b^{-\alpha})/(-\alpha)}\right)^{\frac{1}{\alpha}} & \text{for } \alpha \neq 0,\ \beta = 0 \\ \left(\dfrac{(a^{\beta}-b^{\beta})/\beta}{\log(a/b))}\right)^{\frac{1}{\beta}} & \text{for } \alpha = 0,\ \beta \neq 0 \\ \exp\left(\dfrac{a^{\alpha}\log(eb^{\alpha}) - b^{\alpha}\log(ea^{\alpha})}{\alpha(a^{\alpha}-b^{\alpha})}\right) & \text{for } \alpha = -\beta \neq 0 \\ \sqrt{a\ b} & \text{for } \alpha = \beta = 0 \end{cases} \tag{6.60}$$

determines the value of the constant $\kappa = \mathcal{K}(a,b) \in \mathbb{R}$ that equalizes the value of the AB Log-Det divergences between any arbitrary $a, b \in \mathbb{R}$ constants (in the first argument) and κ (in the second argument), i.e.,

$$D_{AB}^{(\alpha,\beta)}(a\|\ \kappa) = D_{AB}^{(\alpha,\beta)}(b\|\ \kappa). \tag{6.61}$$

Note that the only role of the scaling factor κ is to adjust the reference value in one of the arguments of the divergence to ensure the exact balance in the number of spatial filters that are specialized in each class. As it is shown in the Appendix B, this scaling factor prevents that the minimax solution for $i = 1, \ldots, p$, could be attained by some eigenvectors associated with elements of the inner set of eigenvalues in Eqn. 6.51, so the chosen subset of eigenvectors have to essentially coincide with the principal and minor eigenvectors that form the CSP solution in Eqn. 6.56. In practice, a value of κ which is close to unity and meets the required bounds can be obtained from the truncated choice

$$\kappa_{\star} = \begin{cases} \kappa_{\inf} + \varepsilon & \text{for } \kappa_{\inf} \geq 1 \\ 1 & \text{for } 1 \in (\kappa_{\inf}, \kappa_{sup}) \\ \kappa_{\sup} - \varepsilon & \text{for } \kappa_{\sup} \leq 1 \end{cases} \tag{6.62}$$

for an arbitrary small value of the constant $\epsilon \ll \kappa_{sup} - \kappa_{inf}$.

6.6 The Gradient of the AB Log-Det Divergence

The AB Log-Det divergence between the conditional covariance of the outputs $\mathbf{Y} = \mathbf{W}^T\boldsymbol{x}$ for each of the classes

$$f(\mathbf{W}) = D_{AB}^{(\alpha,\beta)}(Cov(\mathbf{Y}|c_1) \parallel Cov(\mathbf{Y}|c_2)) = D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\|\mathbf{W}^T\mathbf{Q}\mathbf{W}), \quad (6.63)$$

is a function of the matrix $\mathbf{W} \in \mathbb{R}^{n\times p}$.

The optimization of this function with respect to $\mathbf{W}$ is non-trivial, so in this section, the derivation of the gradient of the AB Log-Det divergences is shown. One may note that this is not only naturally interesting for the optimization that we would like to perform in this work, but it also contributes to pave the way for the potential practical use of the AB Log-Det divergence in other scenarios and applications.

As shown previously, the divergence is separable

$$D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\|\mathbf{W}^T\mathbf{Q}\mathbf{W}) = D_{AB}^{(\alpha,\beta)}(\mathbf{M}\|\mathbf{I}_p) = \sum_{i=1}^{p} D_{AB}^{(\alpha,\beta)}(\mu_i(\mathbf{W})\|\, 1) \quad (6.64)$$

over the eigenvalues of the matrix

$$\mathbf{M} = (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{W}^T\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} = \mathbf{U}\,\mathrm{diag}\{\mu_1,\cdots,\mu_p\}\mathbf{U}^T \quad (6.65)$$

where $\mathrm{diag}\{\mu_1,\cdots,\mu_p\}$ and $\mathbf{U} = [\boldsymbol{u}_1,\cdots,\boldsymbol{u}_p]$, respectively denote the matrices of eigenvalues and eigenvectors of $\mathbf{M}$, which are functions of the matrix $\mathbf{W}$. The differential of $f(\mathbf{W})$ can be expressed as

$$df(\mathbf{W}) = \mathrm{tr}\left\{d\mathbf{W}^T\,\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}}\right\} \quad (6.66)$$

where

$$\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} = \left[\frac{\partial f(\mathbf{W})}{\partial W_{ij}}\right]_{ij} \in \mathbb{R}^{n\times p} \quad (6.67)$$

denotes the gradient of the function. The divergence directly depends on the generalized eigenvalues, which in turn depend on the matrix $\mathbf{W}$. The suitable tool to obtain the gradient of this composition of functions is the chain rule, which can be written as

$$\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} = \sum_{i=1}^{p}\frac{\partial \mu_i}{\partial \mathbf{W}}\frac{\partial f(\mathbf{W})}{\partial \mu_i}. \quad (6.68)$$

So, the gradient can be evaluated after finding $\frac{\partial f(\mathbf{W})}{\partial \mu_i}$ and $\frac{\partial \mu_i}{\partial \mathbf{W}}$.

Since the divergence is a separable function of the generalized eigenvalues, the first

term is easier to obtain,

$$\frac{\partial f(\mathbf{W})}{\partial \mu_i} = \frac{\partial D_{AB}^{(\alpha,\beta)}(\mu_i \| \, 1)}{\partial \mu_i} = \begin{cases} \dfrac{\mu_i^{\beta-1} - \mu_i^{-\alpha-1}}{\alpha\mu_i^{\beta} + \beta\mu_i^{-\alpha}} = \dfrac{\mu_i^{\alpha+\beta} - 1}{\mu_i(\alpha\mu_i^{\alpha+\beta} + \beta)} & \text{for } \alpha + \beta \neq 0 \\ \\ \dfrac{\log \mu_i}{\mu_i(1 + \alpha \log \mu_i)} & \text{for } \alpha + \beta = 0 \, . \end{cases} \tag{6.69}$$

Obtaining the second term $\frac{\partial \mu_i}{\partial \mathbf{W}}$ is not so easy and requires to employ our previous plausible assumption that the generalized eigenvalues have a simple spectrum. Under this condition, the *Hadamard first variation formula* can be used to write the differential of the eigenvalues as

$$d\mu_i = \boldsymbol{u}_i^T d\mathbf{M} \boldsymbol{u}_i, \tag{6.70}$$

where $\boldsymbol{u}_i$ denotes the normalized eigenvector ($\|\boldsymbol{u}_i\|_2 = 1$) corresponding to each eigenvalue μ_i.

With the help of the product rule for differentials, we obtain

$$\begin{aligned} d\mathbf{M} = {} & d(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \, (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\mathbf{M} + \mathbf{M}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}} \, d(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \\ & + (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \, (d\mathbf{W}^T\mathbf{P}\mathbf{W} + \mathbf{W}^T\mathbf{P}d\mathbf{W}) \, (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}. \end{aligned} \tag{6.71}$$

As we show in the Appendix C, it can be simplified as follows

$$d(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \, (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}} = -\frac{1}{2}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \, d(\mathbf{W}^T\mathbf{Q}\mathbf{W}) \, (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \tag{6.72}$$

$$= -\frac{1}{2}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \, (d\mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{Q}d\mathbf{W}) \, (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \tag{6.73}$$

hence

$$\begin{aligned} d\mathbf{M} = {} & -\frac{1}{2}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}(d\mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{Q}d\mathbf{W})(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \, \mathbf{M} \\ & - \frac{1}{2}\left[(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}(d\mathbf{W}^T\mathbf{Q}\mathbf{W} + \mathbf{W}^T\mathbf{Q}d\mathbf{W})(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \, \mathbf{M}\right]^T \\ & + (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \, (d\mathbf{W}^T\mathbf{P}\mathbf{W} + \mathbf{W}^T\mathbf{P}d\mathbf{W}) \, (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}. \end{aligned} \tag{6.74}$$

Thus, after substituting Eqn. 6.74 in Eqn. 6.70 and using the invariance of the trace under transpositions ($\mathrm{tr}\{\mathbf{A}\} = \mathrm{tr}\{\mathbf{A}^T\}$) and the cyclic shifts ($\mathrm{tr}\{\mathbf{AB}\} = \mathrm{tr}\{\mathbf{BA}\}$), the

following values are obtained

$$\begin{aligned} d\mu_i &= \boldsymbol{u}_i^T d\mathbf{M}\boldsymbol{u}_i \\ &= \operatorname{tr}\left\{\boldsymbol{u}_i^T d\mathbf{M}\ \boldsymbol{u}_i\right\} \\ &= -\frac{1}{2}\operatorname{tr}\left\{\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}(d\mathbf{W}^T\mathbf{Q}\mathbf{W}+\mathbf{W}^T\mathbf{Q}d\mathbf{W})(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ \mathbf{M}\boldsymbol{u}_i\right\} \\ &\quad -\frac{1}{2}\operatorname{tr}\left\{\boldsymbol{u}_i^T\left[(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}(d\mathbf{W}^T\mathbf{Q}\mathbf{W}+\mathbf{W}^T\mathbf{Q}d\mathbf{W})(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ \mathbf{M}\right]^T\boldsymbol{u}_i\right\} \\ &\quad +\operatorname{tr}\left\{\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\ (d\mathbf{W}^T\mathbf{P}\mathbf{W}+\mathbf{W}^T\mathbf{P}d\mathbf{W})\ (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\boldsymbol{u}_i\right\} \\ &= -2\operatorname{tr}\left\{d\mathbf{W}^T\ \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{M}\ \boldsymbol{u}_i\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\right\} \\ &\quad +2\operatorname{tr}\left\{d\mathbf{W}^T\,\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\boldsymbol{u}_i\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\right\}. \end{aligned} \tag{6.75}$$

At this point, the identity for the differential can be used

$$d\mu_i = \operatorname{tr}\left\{d\mathbf{W}^T\,\frac{\partial\mu_i}{\partial\mathbf{W}}\right\} \tag{6.76}$$

in Eqn. 6.75 to identify the second desired term

$$\begin{aligned} \frac{\partial\mu_i}{\partial\mathbf{W}} &= -2\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{M}\ \boldsymbol{u}_i\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \\ &\quad +2\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\boldsymbol{u}_i\boldsymbol{u}_i^T(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}. \end{aligned} \tag{6.77}$$

Substituting the expressions Eqn. 6.69 and Eqn. 6.77 in Eqn. 6.68, we obtain

$$\begin{aligned} \frac{\partial f(\mathbf{W})}{\partial\mathbf{W}} &= \sum_{i=1}^{p}\frac{\partial\mu_i}{\partial\mathbf{W}}\frac{\partial D_{AB}^{(\alpha,\beta)}(\mu_i\|\,1)}{\partial\mu_i} \\ &= -2\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{M}\ \mathbf{Z}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}+2\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{Z}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} \end{aligned} \tag{6.78}$$

where, for convenience, the matrix is defined as following

$$\mathbf{Z} = \sum_{i=1}^{p}\boldsymbol{u}_i\frac{\partial D_{AB}^{(\alpha,\beta)}(\mu_i\|\,1)}{\partial\mu_i}\boldsymbol{u}_i^T = \mathbf{U}\operatorname{diag}\left\{\frac{\partial D_{AB}^{(\alpha,\beta)}(\mu_1\|\,1)}{\partial\mu_1},\ldots,\frac{\partial D_{AB}^{(\alpha,\beta)}(\mu_p\|\,1)}{\partial\mu_p}\right\}\mathbf{U}^T. \tag{6.79}$$

The matrix $\mathbf{Z}$ can also be represented directly in terms of the matrix $\mathbf{M}$ (which we have defined previously in Eqn. 6.65) as

$$\mathbf{Z} = \begin{cases} \mathbf{M}^{-1}(\alpha\mathbf{M}^{\alpha+\beta}+\beta\mathbf{I})^{-1}(\mathbf{M}^{\alpha+\beta}-\mathbf{I}) & \text{for } \alpha+\beta\neq 0 \\ \\ \mathbf{M}^{-1}((\log\mathbf{M})^{-1}+\alpha\mathbf{I})^{-1} & \text{for } \alpha+\beta=0 \end{cases} \tag{6.80}$$

where $\log(\cdot)$ for matrix arguments denotes the matrix logarithm functional. After the grouping of common terms in Eqn. 6.78 the final gradient expression is obtained, which is given by

$$\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} = 2[\mathbf{P}\mathbf{W} - \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{P}\mathbf{W})](\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{Z}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}. \tag{6.81}$$

6.6.1 Validation of Eqn. 6.81 with the Gradient of the KL Divergence

The Kullback–Leibler (KL) divergence between the Gaussian densities $p(\boldsymbol{x}|c_2)$ and the $p(\boldsymbol{x}|c_1)$, of zero mean and the respective covariance matrices $Cov(\mathbf{Y}|c_1)$ and $Cov(\mathbf{Y}|c_2)$, is given by

$$\begin{aligned} Div_{KL}(p(\boldsymbol{x}|c_2) \parallel p(\boldsymbol{x}|c_1)) = & \int p(\boldsymbol{x}|c_2)\log\frac{p(\boldsymbol{x}|c_2)}{p(\boldsymbol{x}|c_1)}d\boldsymbol{x} && (6.82)\\ = & \ \tfrac{1}{2}\log|Cov(\mathbf{Y}|c_1)| - \tfrac{1}{2}\log|Cov(\mathbf{Y}|c_2)| && (6.83)\\ & + \tfrac{1}{2}\operatorname{tr}\{Cov^{-1}(\mathbf{Y}|c_1)Cov(\mathbf{Y}|c_2) - \mathbf{I}_p\}. \end{aligned}$$

Since this divergence only involves trace and log-det operators, as it is shown in the Appendix D, its gradient with respect to $\mathbf{W}$, i.e.,

$$\begin{aligned} \tfrac{\partial}{\partial \mathbf{W}} Div_{KL}(p(\boldsymbol{x}|c_2) \parallel p(\boldsymbol{x}|c_1)) = & -\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} + \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \\ & + \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} - \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \\ & (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}, \end{aligned} \tag{6.84}$$

is relatively easy to obtain. Then, the fact that the KL divergence is proportional to the AB Log-Det divergence between the class conditional covariance matrices can be used, as long as the conditional covariance matrices appear in the AB Log-Det divergence interchanged in position with respect to class conditional density arguments of the KL divergence. So for the specific case of $\alpha = 1$ and $\beta = 0$, i.e.,

$$D_{AB}^{(1,0)}(Cov(\mathbf{Y}|c_1) \parallel Cov(\mathbf{Y}|c_2)) = 2\, Div_{KL}(p(\boldsymbol{x}|c_2) \parallel p(\boldsymbol{x}|c_1)), \tag{6.85}$$

to test whether there is coherence between the obtained gradient formula in Eqn. 6.81 and twice the gradient of the KL divergence that was independently obtained in the Appendix D. For this purpose, in the specific case of $\alpha = 1$ and $\beta = 0$, from Eqn. 6.80

the following auxiliary matrices are evaluated

$$\mathbf{Z} = \mathbf{M}^{-1}(\mathbf{I}_p - \mathbf{M}^{-1}) \tag{6.86}$$

$$(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{Z}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} = (\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} - (\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \tag{6.87}$$

and are substituted in the expression of the gradient of the AB Log-Det divergence Eqn. 6.81. After the following straightforward simplifications,

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{W}} D_{AB}^{(1,0)}(Cov(\mathbf{Y}|c_1) \,\|\, Cov(\mathbf{Y}|c_2)) &= 2[\mathbf{P}\mathbf{W} - \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{P}\mathbf{W})] \\
&\quad \times [(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{Z}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}] \qquad (6.88)\\
&= 2[-\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{P}\mathbf{W}) + \mathbf{P}\mathbf{W}] \\
&\quad \times [(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} - (\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W}) \\
&\quad (\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}] \qquad (6.89)\\
&= 2[-\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} + \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \\
&\quad + \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} - \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \\
&\quad (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}] \qquad (6.90)\\
&= 2\,\frac{\partial}{\partial \mathbf{W}} Div_{KL}(p(\boldsymbol{x}|c_2) \,\|\, p(\boldsymbol{x}|c_1))\,. \qquad (6.91)
\end{aligned}$$

the proportionality between the gradient of $D_{AB}^{(1,0)}(Cov(\mathbf{Y}|c_1)\|Cov(\mathbf{Y}|c_2))$ and the gradient of the KL divergence in Eqn. 6.84 is confirmed.

6.6.2 Validation of Eqn. 6.81 with the Gradient of the AG Divergence

The Alpha–Gamma divergence between the Gaussian densities $p(\boldsymbol{x}|c_2)$ and $p(\boldsymbol{x}|c_1)$, of zero mean and with respective covariance matrices $Cov(\mathbf{Y}|c_1) = \mathbf{W}^T\mathbf{P}\mathbf{W}$ and $Cov(\mathbf{Y}|c_2) = \mathbf{W}^T\mathbf{Q}\mathbf{W}$, is equal to

$$Div_{AG}^{(\alpha,\beta)}\left(p(y_i|c_2)\|p(y_i|c_1)\right) \equiv \frac{1}{\alpha\beta}\log\frac{\left(\int_\Omega p(y_i|c_1)\,dy_i\right)^\beta \left(\int p(y_i|c_2)\,dy_i\right)^\alpha}{\int p^\beta(y_i|c_1)\,p^\alpha(y_i|c_2)\,dy_i} \tag{6.92}$$

$$\begin{aligned}
&= \frac{1}{2\alpha\beta}\log|\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W}| - \frac{1}{2\beta}\log|\mathbf{W}^T\mathbf{P}\mathbf{W}| \\
&\quad - \frac{1}{2\alpha}\log|\mathbf{W}^T\mathbf{Q}\mathbf{W}| \\
&\quad \text{for} \quad \alpha > 0,\ \beta > 0,\ \alpha + \beta = 1. \qquad (6.93)
\end{aligned}$$

Due to the constraint $\alpha + \beta = 1$, it is assume that β is determined by α, i.e., $\beta = 1 - \alpha$ along this subsection. Since

$$\nabla_{\mathbf{W}} \log|(\mathbf{W}^T\mathbf{P}\mathbf{W})| = 2\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}\,, \tag{6.94}$$

the gradient of the AG divergence with respect to $\mathbf{W}$ is given by

$$\begin{aligned}\frac{\partial}{\partial \mathbf{W}} Div_{AG}^{(\alpha,\beta)}(p(\boldsymbol{x}|c_2) \,\|\, p(\boldsymbol{x}|c_1)) &= \tfrac{2}{2\alpha\beta}(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W}\left(\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W}\right)^{-1} \\ &\quad - \tfrac{2}{2\alpha}\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} - \tfrac{2}{2\beta}\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \\ &= -\tfrac{1}{\beta}\mathbf{P}\mathbf{W}[(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} - \left(\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W}\right)^{-1}] \\ &\quad - \tfrac{1}{\alpha}\mathbf{Q}\mathbf{W}[(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} - \left(\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W}\right)^{-1}].\end{aligned} \tag{6.95}$$

Then, the equivalence between the AG divergence and the AB Log-Det divergence between the class conditional covariance matrices can be used

$$D_{AB}^{(\alpha,\beta)}(Cov(\mathbf{Y}|c_1) \,\|\, Cov(\mathbf{Y}|c_2)) = 2\, Div_{AG}^{(\alpha,\beta)}\left(p(y_i|c_2)\|p(y_i|c_1)\right), \tag{6.96}$$

which is valid for the specific case of $\alpha + \beta = 1$ and $\alpha, \beta > 0$, to also test the coherence between the obtained gradient formula in Eqn. 6.81 and twice the gradient of the AG divergence. For $\alpha + \beta = 1$, the auxiliary matrices in the definition of the gradient are

$$\mathbf{Z} = (\alpha\mathbf{M} + \beta\mathbf{I})^{-1}[\mathbf{M}^{-1}(\mathbf{M} - \mathbf{I})] = (\alpha\mathbf{M} + \beta\mathbf{I})^{-1} - (\alpha\mathbf{M}^2 + \beta\mathbf{M})^{-1} \tag{6.97}$$

and

$$\begin{aligned}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}}\mathbf{Z}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-\frac{1}{2}} &= (\alpha(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\mathbf{M}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}} + \beta\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} \\ &\quad - (\alpha(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\mathbf{M}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} \\ &\quad (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\mathbf{M}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}} + \beta(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}}\mathbf{M} \\ &\quad (\mathbf{W}^T\mathbf{Q}\mathbf{W})^{\frac{1}{2}})^{-1} \qquad (6.98) \\ &= [\mathbf{I} - (\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})](\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W})^{-1}\,. \qquad (6.99)\end{aligned}$$

After substituting this last expression in the gradient of the AB Log-Det divergence Eqn. 6.81, we obtain

$$\frac{\partial}{\partial \mathbf{W}} D_{AB}^{(\alpha,\beta)}(Cov(\mathbf{Y}|c_1) \,\|\, Cov(\mathbf{Y}|c_2))$$

$$= 2[\mathbf{PW} - \mathbf{QW}(\mathbf{W}^T\mathbf{QW})^{-1}(\mathbf{W}^T\mathbf{PW})]$$
$$(\mathbf{W}^T\mathbf{QW})^{-\frac{1}{2}}\mathbf{Z}(\mathbf{W}^T\mathbf{QW})^{-\frac{1}{2}} \tag{6.100}$$
$$= +2(\mathbf{PW} + \mathbf{QW})(\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W})^{-1}$$
$$- \tfrac{2}{\beta}\alpha\mathbf{PW}[(\mathbf{W}^T\alpha\mathbf{PW}) + (\mathbf{W}^T\alpha\mathbf{PW})$$
$$(\mathbf{W}^T\beta\mathbf{QW})^{-1}(\mathbf{W}^T\alpha\mathbf{PW})]^{-1} - \tfrac{2}{\alpha}\beta\mathbf{QW}[(\mathbf{W}^T\beta\mathbf{QW})$$
$$+ (\mathbf{W}^T\beta\mathbf{QW})(\mathbf{W}^T\alpha\mathbf{PW})^{-1}(\mathbf{W}^T\beta\mathbf{QW})]^{-1}\,. \tag{6.101}$$

With the help of the particular form of the Woodbury identity for the matrix inverse

$$[\mathbf{A} + \mathbf{AB}^{-1}\mathbf{A}]^{-1} = \mathbf{A}^{-1} - (\mathbf{A} + \mathbf{B})^{-1} \tag{6.102}$$

we simplify the terms within the brackets. Finally, the fact that $\alpha + \beta = 1$ is used to confirm the proportionality with the gradient of the AG divergence given in Eqn. 6.95,

$$\tfrac{\partial}{\partial \mathbf{W}} D_{AB}^{(\alpha,\beta)}(Cov(\mathbf{Y}|c_1) \,\|\, Cov(\mathbf{Y}|c_2)) = +2(\mathbf{PW} + \mathbf{QW})(\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W})^{-1}$$
$$- \tfrac{2}{\beta}\alpha\mathbf{PW}[(\mathbf{W}^T\alpha\mathbf{PW})^{-1} - (\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W})^{-1}]$$
$$- \tfrac{2}{\alpha}\beta\mathbf{QW}[(\mathbf{W}^T\beta\mathbf{QW})^{-1} - (\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W})^{-1}] \tag{6.103}$$
$$= +2(\mathbf{PW} + \mathbf{QW})(\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W})^{-1}$$
$$- \tfrac{2}{\beta}\mathbf{PW}[(\mathbf{W}^T\mathbf{PW})^{-1} - \alpha(\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W})^{-1}]$$
$$- \tfrac{2}{\alpha}\mathbf{QW}[(\mathbf{W}^T\mathbf{QW})^{-1} - \beta(\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W})^{-1}] \tag{6.104}$$
$$= +2((1 + \tfrac{\alpha}{\beta})\mathbf{PW} + (1 + \tfrac{\beta}{\alpha})\mathbf{QW})(\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})$$
$$\mathbf{W})^{-1} - \tfrac{2}{\beta}\mathbf{PW}(\mathbf{W}^T\mathbf{PW})^{-1} - \tfrac{2}{\alpha}\mathbf{QW}(\mathbf{W}^T\mathbf{QW})^{-1} \tag{6.105}$$
$$= +2(\tfrac{1}{\beta}\mathbf{PW} + \tfrac{1}{\alpha}\mathbf{QW})(\mathbf{W}^T(\alpha\mathbf{P} + \beta\mathbf{Q})\mathbf{W})^{-1}$$
$$- \tfrac{2}{\beta}\mathbf{PW}(\mathbf{W}^T\mathbf{PW})^{-1} - \frac{2}{\alpha}\mathbf{QW}(\mathbf{W}^T\mathbf{QW})^{-1} \tag{6.106}$$
$$= 2\tfrac{\partial}{\partial \mathbf{W}} Div_{AG}^{(\alpha,\beta)}(p(\boldsymbol{x}|c_2) \,\|\, p(\boldsymbol{x}|c_1))\,. \tag{6.107}$$

6.7 Robustness of the AB Log-Det Divergence in Terms of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

The squared Riemann metric is known to be the natural distance in the manifold of SPD matrices, as it measures the squared length of the geodesic path between the arguments of the divergence (Harandi et al., 2017). However, in the real data there are usually several model contaminations (mismatches), including outliers or artifacts, that could make other robust divergences preferable. In this section, we study how the hyperparameters α and β can influence robustness of the AB Log-Det divergence with respect to the behavior of the squared Riemann metric, which is used as a reference.

For convenience, the AB Log-Det divergence is denoted as a function of the spatial filter matrix $\mathbf{W}$ by

$$f_{(\alpha,\beta)}(\mathbf{W}) \equiv D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\|\mathbf{W}^T\mathbf{Q}\mathbf{W}), \tag{6.108}$$

and its gradient expression given by Eqn. 6.68 is considered. The spatial filters that maximize this divergence should satisfy the following estimating equations

$$\frac{\partial f_{(\alpha,\beta)}(\mathbf{W})}{\partial \mathbf{W}} = \sum_{i=1}^{p} \frac{\partial \mu_i}{\partial \mathbf{W}} \psi_{(\alpha,\beta)}(\mu_i) = 0, \tag{6.109}$$

where $\mu_i,\ i = 1, \ldots, p$, are the eigenvalues of matrix $\mathbf{M}$, which was defined in Eqn. 6.65, and

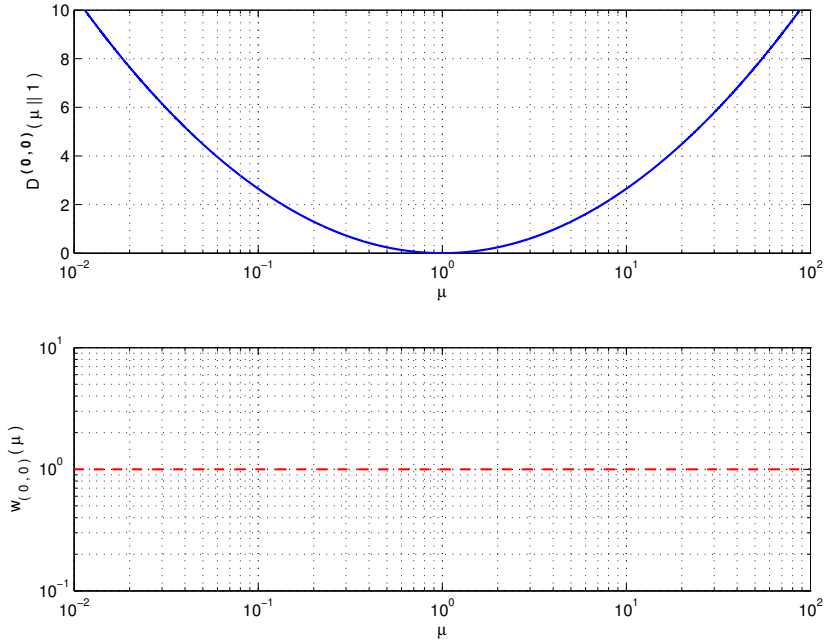
$$\psi_{(\alpha,\beta)}(\mu_i) = \frac{\partial f_{(\alpha,\beta)}(\mathbf{W})}{\partial \mu_i}, \quad i = 1, \ldots, p, \tag{6.110}$$

may be regarded as influence functions for each pair (α, β) that account for the penalty variation in the divergence with respect to μ_i. The complementary term to $\psi_{(\alpha,\beta)}(\mu_i)$ in Eqn. 6.109, i.e., $\frac{\partial \mu_i}{\partial \mathbf{W}}$, is a matrix of partial derivatives of the generalized eigenvalues μ_i with respect to the elements of the spatial filters $\mathbf{W}$ and, therefore, it is independent of the considered divergence. It is easy to observe that, in the particular case of $\alpha = \beta = 0$, the expression in Eqn. 6.109 represents the estimating equation for the squared Riemann metric

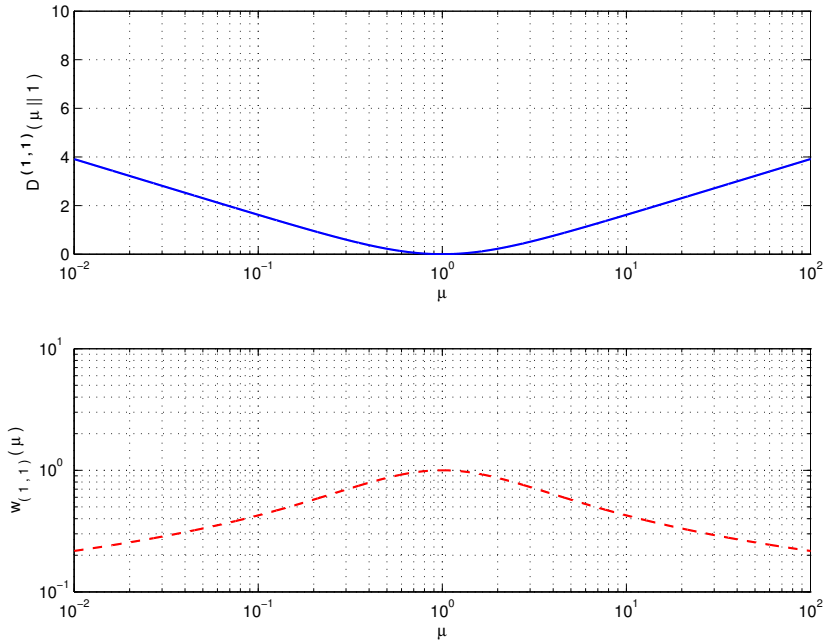
$$\frac{\partial f_{(0,0)}(\mathbf{W})}{\partial \mathbf{W}} = \sum_{i=1}^{p} \frac{\partial \mu_i}{\partial \mathbf{W}} \psi_{(0,0)}(\mu_i) = 0. \tag{6.111}$$

In order to study the relative robusness to outliers, one can rewrite the estimating equation for a chosen pair of hyperparameters (α, β) in terms of the influence function for the squared Riemannian metric as

$$\frac{\partial f_{(\alpha,\beta)}(\mathbf{W})}{\partial \mathbf{W}} = \sum_{i=1}^{p} \left(\frac{\partial \mu_i}{\partial \mathbf{W}} \psi_{(0,0)}(\mu_i) \right) w_{(\alpha,\beta)}(\mu_i) = 0, \tag{6.112}$$

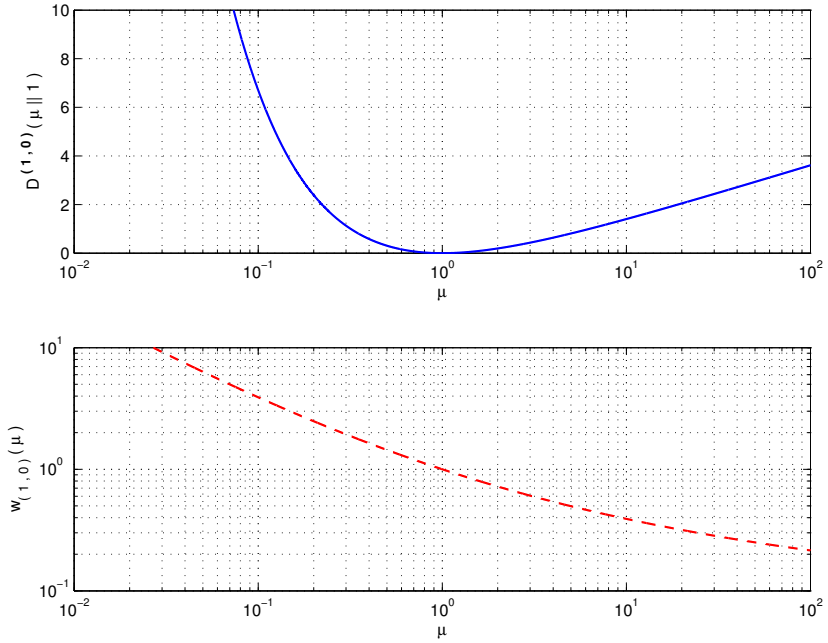


(a)

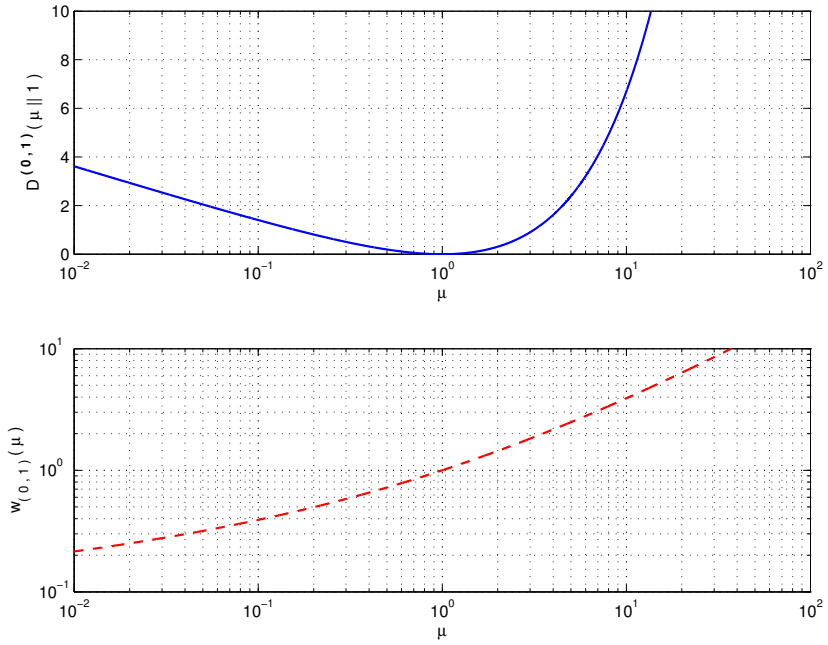


(b)

Fig. 6.3 Illustration of the behavior of the AB Log-Det divergence $D_{AB}^{(\alpha,\beta)}(\mu, 1)$, and of its associated weight function $w_{\alpha,\beta}(\mu)$, versus μ for different values of $\alpha = \beta$. Note that μ is shown in log-scale. (a) Squared Riemannian metric for $\alpha = \beta = 0$ (upper plot) and its weight function (lower plot); (b) Power Log-Det divergence for $\alpha = \beta = 1$ (upper plot) and its weight function (lower plot).



(a)



(b)

Fig. 6.4 Illustration of the behavior of the AB Log-Det divergence $D_{AB}^{(\alpha,\beta)}(\mu, 1)$, and of its associated weight function $w_{\alpha,\beta}(\mu)$, versus μ for different values of $\alpha \neq \beta$. Note that μ is shown in log-scale. (a) Kullback–Leibler (KL) positive definite matrix divergence for $\alpha = 1, \beta = 0$, and its weight function (lower plot); (b) Dual KL positive definite matrix div. for $\alpha = 0, \beta = 1$, and its weight function (lower plot).

where the scalar term

$$w_{(\alpha,\beta)}(\mu) = \frac{\psi_{(\alpha,\beta)}(\mu)}{\psi_{(0,0)}(\mu)} \tag{6.113}$$

acts as a weight function that controls, for a given pair (α, β), the magnitude of the effect in the estimation equation of departures of μ_i from unity.

The presence of outliers in the real data, typically results in eigenvalues μ_i that are too far from unity. However, depending on the problem, the higher prevalence of outliers may be stronger only for the greatest eigenvalues, or for the smallest eigenvalues, or simultaneously for the greatest and smaller eigenvalues. Those hyperparameters (α, β) that are able to down-weight the contribution of the outliers, are considered more robust. Therefore, the shape of the weight functions $w_{(\alpha,\beta)}(\mu_i)$ is useful to study the relative immunity of the AB Log-Det divergence to outliers.

Fig. 6.3a shows the squared Riemannian metric ($\alpha = \beta = 0$) and its weight function, which is flat since this divergence is taken as reference. Fig. 6.3b presents a similar plot for the Power Log-det divergence with $\alpha = \beta = 1$. In this case, the bell shape of the weight function is an indicator of the robustness with respect to the presence of outliers in the greatest and smallest eigenvalues, since they will be down-weighted in the estimating Eqn. 6.112. Similar plots can be done by increasing the magnitude of $\alpha = \beta$, which progressively enhances the robustness. When $\alpha \neq \beta$ the divergence is asymmetric. Fig. 6.4 respectively present the Kullback–Leibler divergence for SPD matrices ($\alpha = 1, \beta = 0$) and its dual version ($\alpha = 0, \beta = 1$), together with their associated weight functions. These plots illustrate the asymmetric cases in situations where $\alpha + \beta > 0$ and reveal that, when $\alpha \gg \beta$, the AB Log-Det divergences tend to be more robust against outliers in the large eigenvalues while, for $\alpha \ll \beta$, the robustness tends to be with respect to the outliers in small eigenvalues.

6.8 Conclusions

The key properties of AB Log-Det divergences have been summarized. We have reexamined the relation between the Common Spatial Pattern criterion with a predefined number of spatial filters for each class and its interpretation as an AB Log-Det divergence optimization problem, to show that a scaling factor in one of the arguments is necessary for the equivalence of the solutions.

CHAPTER 7

Optimization of Alpha-Beta Log Det Divergence Algorithm

It is mentioned in Chapter 6 that the optimization of AB Log-Det divergence is non-trivial. This motivates us to use these divergences with the illustrative application of dimensionality reduction in BCI and explain how to perform their optimization. The EEG data has a typical high-dimensionality, a low signal to noise ratio and may have artifacts/outliers. The dimensionality reduction is then a necessary processing of the EEG signals for extracting those subspaces where the features have highest discriminative power.

In this chapter, the Sub-ABLD algorithm is proposed based on the theoretical background presented in Chapter 6 for the discrimination of two class motor imagery movements. The chapter starts with the related reviews of the spatial filtering for MI movements. The proposed criterion and algorithm are presented in Section 7.2 and Section 7.3 describes the experimental study. The results obtained are presented in Section 7.4.

7.1 Review of Some Related Techniques for the Spatial Filtering of Motor Imagery Movements

It is well known that the performance of CSP is easily affected by the presence of artifacts. To overcome this drawback several CSP variants algorithm have been proposed which has been discussed in Chapter 3. In this section, the related regularized variants of CSP that have been proposed to improve the classification performance is reviewed. The regularization approaches of CSP are mainly done either in the estimation of the covariance matrices or by modifying the CSP objective function.

Most of them combine the estimation of the covariance matrices for each class with the regularization of the CSP objective function using penalty terms. Some of the approaches include the previous information (Lotte and Guan, 2010*b*), other subject data (Kang et al., 2009; Lotte and Guan, 2010*a*) and previous session data (Lu et al., 2009) for estimating the class covariance matrix. Another approach used M-estimators to compute the robust class covariance matrices (Xinyi Yong and Birch, 2008) and yet another approach obtained the covariance matrices by finding the minimum squared error (Kawanabe and Vidaurre, 2009). The authors of (Samek, Binder and Müller, 2013)

applied Multiple Kernel Learning (MKL) to combine the information from different subjects.

It has been shown in (Lotte and Guan, 2011) that the regularization of the objective function is more useful than regularizing the estimated covariance matrix. Several approaches have been proposed by regularizing the objective function. The authors of (Blankertz et al., 2007) have additionally incorporated the EOG signals for reducing the ocular artifacts. Other authors have tried to ensure robustness by selecting only the important channels and produce sparse spatial filters (Arvaneh et al., 2011*b*; Farquhar et al., 2006; Yong et al., 2008*b*). Another approach is to robustify the system by obtaining only the stationary features. A robustify maximin CSP method was proposed that used a set of covariance matrices instead of an individual covariance matrix without using any other user data or data from the previous sessions (Kawanabe et al., 2009, 2014). In order to avoid the presence of the outlier, the CSP objective function has been formulated using l_p-norm in (Wang et al., 2012; Park and Chung, 2013). The Stationary Subspace Analysis (SSA) algorithm was proposed to obtain the stationary subspaces of the time series EEG signals by considering only the stationary components of the signals. The limitation of this method is the detection of dissimilarity of the different class as a non-stationary feature (Von Bünau et al., 2009). The group wise SSA (gwSSA) algorithm aims at obtaining the non-stationarities by dividing the dataset into different groups and calculating the minimum KL divergence between estimated source distribution of each trial in a group and the average distribution of the corresponding group. This algorithm not only allows the combining of the multisubject data but also the multiclass data (Samek et al., 2011). But, the gwSSA algorithm cannot find the discriminative information between the classes. The same group proposed a new approach for extracting the discriminative information, by subtracting the inter class divergences from the gwSSA objective function (Samek, Müller, Kawanabe and Vidaurre, 2012). To overcome the limitation of the SSA algorithm, two-step approaches have been proposed where the initial extraction of the stationary sources was done using the SSA method and later, the CSP was used for the computation of the spatial filters(Von Bünau et al., 2010). Another approach to extract the stationary features is to reduce the nonstationarities between the two sessions. The supervised and unsupervised methods for adaptation of the data space have been proposed using KL divergence between the intersession data (Arvaneh et al., 2013). Recently, the authors of (Wu et al., 2015) presented MAP-CSP algorithm by deriving the probabilistic model of CSP to resolve the issue of overfitting of the baseline CSP algorithm.

One of the limitations of the CSP algorithm is that it is mainly suitable only for the discrimination of two classes, while, in general, for an efficient BCI system more than two motor imagery movements are required. In order to formulate it for the multiclass system, the authors of (Müller-Gerking et al., 1999; Allwein et al., 2001) have

reduced the multiclass problem to a binary problem. The authors of (Dornhege et al., 2004) proposed two approaches for the multiclass problem; firstly to find the spatial filters for one class with respect to all the other classes and secondly, by simultaneous diagonalization methods. Other approaches, like (Grosse-Wentrup and Buss, 2008), proposed to solve the multiclass problems by combining information theoretic criteria with joint diagonalization methods. Several other methods have been proposed for the multiclass paradigm using independent component analysis (Naeem et al., 2006) and Riemannian geometry to obtain the spatial filters (Barachant et al., 2012). The authors of (Zhang et al., 2013) derived a relation between Bayes classification error and Rayleigh quotient and used this approach to solve the multiclass problem. In spite of all these different approaches, the performance of MI-based BCI systems is degraded due to the presence of non-stationarities and outliers, which is a challenge for the BCI systems in a real application. Hence, a robust feature extraction algorithm is needed to increase the overall performance of the system.

7.2 Proposed Criterion and Algorithm for Spatial Filtering

For the presentation of the proposed criterion some additional notation needs to be defined. Let $\tilde{\boldsymbol{x}}^{(j)}(t)|c$ denote the output of the passband filtering of the raw observations at time t and for the jth trial of class $c \in \{c_1, c_2\}$. The power of the trials of a given class c is normalized by the operation

$$\boldsymbol{x}^{(j)}(t) = \frac{\tilde{\boldsymbol{x}}^{(j)}(t)}{\sqrt{\mathrm{tr}\{Cov(\tilde{\boldsymbol{x}}^{(j)}|c)\}}}, \tag{7.1}$$

where

$$Cov(\boldsymbol{x}^{(j)}|c) = \frac{1}{L}\sum_{t=1}^{L}\left(\boldsymbol{x}^{(j)}(t) - \bar{\boldsymbol{x}}^{(j)}\right)\left(\boldsymbol{x}^{(j)}(t) - \bar{\boldsymbol{x}}^{(j)}\right)^T \quad \text{with} \quad \bar{\boldsymbol{x}}^{(j)} = \frac{1}{L}\sum_{t=1}^{L}\boldsymbol{x}^{(j)}(t) \tag{7.2}$$

denotes the sample covariance matrix the j^{th} trial $\boldsymbol{x}^{(j)}$ of class c, and L is the size in samples of each trial. In order to simplify the notation, the covariance matrices of the two classes are renamed as

$$\mathbf{P}_j \equiv Cov(\boldsymbol{x}^{(j)}|c_1) \quad \text{and} \quad \mathbf{Q}_j \equiv Cov(\boldsymbol{x}^{(j)}|c_2), \tag{7.3}$$

and their averaged versions (the centroids of each class) are denoted as

$$\mathbf{P} \equiv <\mathbf{P}_j> = \frac{1}{N_1}\sum_{j=1}^{N_1}\mathbf{P}_j \quad \text{and} \quad \mathbf{Q} \equiv <\mathbf{Q}_j> = \frac{1}{N_2}\sum_{j=1}^{N_2}\mathbf{Q}_j. \tag{7.4}$$

The classification of imagery movements involves extracting the relevant features of the observations and the classification of the observed patterns in the feature space. In the considered application, the data is high-dimensional but only a few features are sufficient to capture the discriminative information about the intended movements. Thus, the extraction of the relevant features involves a dimensionality reduction step for the observations from $\mathbb{R}^n$ to $\mathbb{R}^p$ where $p \ll n$. This step is implemented through the spatial filtering, i.e., by projecting the n-dimensional observations onto a p-dimensional subspace which should allow a good discrimination of the cluster centroids and, at the same time, guarantee a compact representation of the clusters.

As mentioned earlier, the CSP solution will be obtained by a minimax optimization of the divergence between the projected and scaled centroids of the classes, i.e., $D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T\mathbf{P}\boldsymbol{w}_i \parallel \kappa\ \boldsymbol{w}_i^T\mathbf{Q}\boldsymbol{w}_i)$. However, since this solution completely ignores the within-class dispersion of the samples, it is quite sensitive to artifact and outlier in the training dataset. In similarity with the divergence framework presented in (Samek et al., 2014) and with some variants of Fisher LDA (Barber, 2012, pag. 366), one can regularize the previous problem by controlling the dispersion of the trials of each class around their centroids and also by exploiting the degrees of freedom in the selection of the hyperparameters of the divergences. Then, a robust criterion based on the AB Log-Det divergence takes the following form

$$F(\mathbf{W}) = D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W} \| \kappa\, \mathbf{W}^T\mathbf{Q}\mathbf{W}) - \eta\left(p(c_1)\mathbf{R}_1 + p(c_2)\mathbf{R}_2\right)\ , \tag{7.5}$$

where the penalties associated to the within-class dispersion involve the averaged divergences

$$\mathbf{R}_1 = \frac{1}{N_1}\sum_{j=1}^{N_1} D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}_j\mathbf{W} \| \mathbf{W}^T\mathbf{P}\mathbf{W}), \tag{7.6}$$

$$\mathbf{R}_2 = \frac{1}{N_2}\sum_{j=1}^{N_2} D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{Q}_j\mathbf{W} \| \mathbf{W}^T\mathbf{Q}\mathbf{W}), \tag{7.7}$$

and the parameter $\eta \in \mathbb{R}^+$ controls the balance between the maximization of the between-class scatter and the minimization of the within-class scatter. Note that in Eqn. 7.7 the fact that the AB Log-Det divergence is invariant under the common scaling of its arguments is used, to simplify $D_{AB}^{(\alpha,\beta)}(\kappa\mathbf{W}^T\mathbf{Q}_j\mathbf{W} \| \kappa\mathbf{W}^T\mathbf{Q}\mathbf{W}) = D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{Q}_j\mathbf{W} \| \mathbf{W}^T\mathbf{Q}\mathbf{W})$.

The optimization of the criterion in Eqn. 7.5 can be performed simultaneously, for all the spatial filters, with the use of subspace techniques (Samek et al., 2014). In the next section, a subspace optimization algorithm based on AB Log-Det divergences is presented.

7.2.1 The Subspace Optimization Algorithm (Sub-ABLD)

The subspace method aims to extract the desired set of p spatial filters in two steps. The idea is to first use a robust method to determine the discriminative subspace of the spatial filters, for instance, considering the optimization of a robust criterion like Eqn. 7.5. Later, another criterion is used to identify the individual spatial filters within the subspace. Since the influence of outliers on the solution is significantly reduced after the discriminative subspace is determined. In the second step, the standard CSP criterion can be safely used to determine the final spatial directions within the chosen subspace.

The input parameters of the subspace optimization algorithm based on AB Log-Det divergences (Sub-ABLD) are the set of covariance matrices for each class ($\mathbf{P}_j$, $\mathbf{Q}_j$), the dimension of subspace to be extracted p, and the hyperparameters α, β and η. The method starts with the computation of the sample prior probabilities as well as the average covariance matrices for each class, i.e., $p(c_1), p(c_2)$ and ($\mathbf{P}$, $\mathbf{Q}$). The spatial filter matrix decomposes as $\mathbf{W}^T = \mathbf{\Omega}^T\mathbf{T}$ into the product of a whitening transformation matrix $\mathbf{T}$ of the observations and a semi-orthogonal matrix $\mathbf{\Omega}^T$, which satisfies $\mathbf{\Omega}^T\mathbf{\Omega} = \mathbf{I}_p$. The whitening transformation is obtained from eigenvalue decomposition of $Cov(\boldsymbol{x}) = p(c_1)\mathbf{P} + p(c_2)\mathbf{Q} = \mathbf{U}_1\mathbf{\Delta}\mathbf{U}_1^T$ as follows

$$\mathbf{T} = \mathbf{\Delta}^{-\frac{1}{2}}\mathbf{U}_1^T, \tag{7.8}$$

where $\mathbf{\Delta}$ and $\mathbf{U}_1$ represent the matrices of eigenvalues and eigenvectors. This transformation is applied to both sides of the covariance matrices to obtain the whitened trial covariances

$$\breve{\mathbf{P}}_j = \mathbf{T}\,\mathbf{P}_j\mathbf{T}^T, \qquad \breve{\mathbf{Q}}_j = \mathbf{T}\,\mathbf{Q}_j\mathbf{T}^T, \tag{7.9}$$

and their averaged versions

$$\breve{\mathbf{P}} = \mathbf{T}\,\mathbf{P}\,\mathbf{T}^T, \qquad \breve{\mathbf{Q}} = \mathbf{T}\,\mathbf{Q}\,\mathbf{T}^T. \tag{7.10}$$

The scaling parameter κ, which pursues the balance of the number of features for each class in absence of regularizers, is determined with the truncation procedure proposed in Eqn. 6.62. The semiorthogonal matrix $\mathbf{\Omega}^T$ that projects the whitened observations onto a p-dimensional subspace is initialized from the identity matrix of dimension $n \times p$. This is equivalent to start the optimization projecting onto the principal p-dimensional subspace of the observations, which ensures a good initial signal to noise ratio. Once the whitening transformation is fixed, the criterion to optimize $F(\mathbf{W})$ can be rewritten, in

terms of $\boldsymbol{\Omega}$, as the following function

$$f(\boldsymbol{\Omega}) = D_{AB}^{(\alpha,\beta)}(\boldsymbol{\Omega}^T\breve{\mathbf{P}}\boldsymbol{\Omega}\|\kappa\boldsymbol{\Omega}^T\breve{\mathbf{Q}}\boldsymbol{\Omega}) - \eta\Bigg((p(c_1)\frac{1}{N_1}\sum_{j=1}^{N_1} D_{AB}^{(\alpha,\beta)}(\boldsymbol{\Omega}^T\breve{\mathbf{P}}_j\boldsymbol{\Omega}\|\boldsymbol{\Omega}^T\breve{\mathbf{P}}\boldsymbol{\Omega}) + p(c_2)\frac{1}{N_2}\sum_{j=1}^{N_2} D_{AB}^{(\alpha,\beta)}(\boldsymbol{\Omega}^T\breve{\mathbf{Q}}_j\boldsymbol{\Omega}\|\boldsymbol{\Omega}^T\breve{\mathbf{Q}}\boldsymbol{\Omega})\Bigg), \tag{7.11}$$

which ordinary gradient can be determined from Eqn. 6.81, to obtain

$$\begin{aligned}\frac{\partial f(\boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}} &= 2[\breve{\mathbf{P}}\boldsymbol{\Omega} - \kappa\breve{\mathbf{Q}}\boldsymbol{\Omega}(\kappa\boldsymbol{\Omega}^T\breve{\mathbf{Q}}\boldsymbol{\Omega})^{-1}(\boldsymbol{\Omega}^T\breve{\mathbf{P}}\boldsymbol{\Omega})](\kappa\boldsymbol{\Omega}^T\breve{\mathbf{Q}}\boldsymbol{\Omega})^{-\frac{1}{2}}\mathbf{Z}_1(\kappa\boldsymbol{\Omega}^T\breve{\mathbf{Q}}\boldsymbol{\Omega})^{-\frac{1}{2}} \\ &- \eta\Bigg((p(c_1)\tfrac{2}{N_1}\sum_{j=1}^{N_1}[\breve{\mathbf{P}}_j\boldsymbol{\Omega} - \breve{\mathbf{P}}\boldsymbol{\Omega}(\boldsymbol{\Omega}^T\breve{\mathbf{P}}\boldsymbol{\Omega})^{-1}(\boldsymbol{\Omega}^T\breve{\mathbf{P}}_j\boldsymbol{\Omega})](\boldsymbol{\Omega}^T\breve{\mathbf{P}}\boldsymbol{\Omega})^{-\frac{1}{2}}\mathbf{Z}_2(\boldsymbol{\Omega}^T\breve{\mathbf{P}}\boldsymbol{\Omega})^{-\frac{1}{2}}\Bigg). \\ &+ p(c_2)\tfrac{2}{N_2}\sum_{j=1}^{N_2}[\breve{\mathbf{Q}}_j\boldsymbol{\Omega} - \breve{\mathbf{Q}}\boldsymbol{\Omega}(\boldsymbol{\Omega}^T\breve{\mathbf{Q}}\boldsymbol{\Omega})^{-1}(\boldsymbol{\Omega}^T\breve{\mathbf{Q}}_j\boldsymbol{\Omega})](\boldsymbol{\Omega}^T\breve{\mathbf{Q}}\boldsymbol{\Omega})^{-\frac{1}{2}}\mathbf{Z}_3(\boldsymbol{\Omega}^T\breve{\mathbf{Q}}\boldsymbol{\Omega})^{-\frac{1}{2}}\Bigg)\end{aligned} \tag{7.12}$$

where the matrices $\mathbf{Z}_i$ should be defined for each case ($i = 1, \ldots, 3$) as in Eqn. 6.80. However, this gradient is not the fastest ascent direction in the structured manifold of semi-orthogonal matrices (the Stiefel manifold). Instead, the fastest ascent direction is given by the "natural" gradient in this manifold (Edelman et al., 1998; Amari, 1998), which is given by

$$\nabla_\Omega f(\boldsymbol{\Omega}) = \frac{\partial f(\boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}} - \boldsymbol{\Omega}\left(\frac{\partial f(\boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}}\right)^T\boldsymbol{\Omega}. \tag{7.13}$$

Let $\boldsymbol{\Omega}^{(i)}$ denote the semi-orthogonal matrix at iteration i and let $\mu^{(i)}$ denotes the step-size, the gradient ascent update is then performed with

$$\boldsymbol{\Omega}_{tg}^{(i+1)} = \boldsymbol{\Omega}^{(i)} + \mu^{(i)}\,\nabla_\Omega f(\boldsymbol{\Omega}^{(i)}). \tag{7.14}$$

The resulting matrix $\boldsymbol{\Omega}_{tg}^{(i+1)}$ belongs to the tangent space of the manifold at $\boldsymbol{\Omega}^{(i)}$ and asymptotically follows the geodesic path of maximum ascent for a sufficient small stepsize $\mu \to 0$. However, for practical stepsizes, like the one that we consider next

$$\mu^{(i)} = \frac{0.02}{\|\nabla_\Omega f(\boldsymbol{\Omega}^{(i)})\|_F}, \tag{7.15}$$

the resulting updates $\boldsymbol{\Omega}_{tg}^{(i+1)}$ are not exactly semi-orthogonal and, in order to restore this property, a retraction procedure onto the manifold is necessary after each iteration. The retraction can be implemented with the help of the MatLab command for a "thin"

singular value decomposition as

$$[\mathbf{Q}_L, \mathbf{D}, \mathbf{Q}_R] = \text{svd}(\mathbf{\Omega}_{tg}^{(i+1)}, 0)\,, \tag{7.16}$$

$$\mathbf{\Omega}^{(i+1)} = \mathbf{Q}_L \mathbf{Q}_R^T\,. \tag{7.17}$$

The procedure is then repeated until convergence to a maxima of the criterion at a given iteration i_{max}. After that, the solution $(\mathbf{\Omega}^{(i_{max})})^T\mathbf{T}$ identifies the subspace of the spatial filters, but not each of their individual directions. In order to determine them, one can solve a CSP problem within the previously identified subspace. We compute the generalized eigenvalues of the matrix pencil $((\mathbf{\Omega}^{(i_{max})})^T\breve{\mathbf{P}}\mathbf{\Omega}^{(i_{max})},\ (\mathbf{\Omega}^{(i_{max})})^T\breve{\mathbf{Q}}\mathbf{\Omega}^{(i_{max})})$ and use the resulting principal and minor eigenvectors $\breve{\boldsymbol{v}}_j$ to form the spatial filter matrix

$$\breve{\mathbf{V}} = [\breve{\boldsymbol{v}}_1, \ldots, \breve{\boldsymbol{v}}_{\lfloor \frac{p}{2} \rfloor}, \breve{\boldsymbol{v}}_{n-p+1+\lfloor \frac{p}{2} \rfloor}, \ldots, \breve{\boldsymbol{v}}_n]. \tag{7.18}$$

The final matrix of spatial filters that solves the problem, is the product of the whitening matrix $\mathbf{T}$, the projection matrix $(\mathbf{\Omega}^{(i_{max})})^T$ and a CSP rotation matrix $\breve{\mathbf{V}}^T$ which operates within the subspace, i.e.,

$$\mathbf{W}^T = \breve{\mathbf{V}}^T(\mathbf{\Omega}^{(i_{max})})^T\mathbf{T}. \tag{7.19}$$

The proposed subspace algorithm (Sub-ABLD) is similar in structure to the one presented in (Samek et al., 2014) for Beta divergences. In spite of the fact that they optimize different criteria, the main difference between both subspace algorithms is in the specific way that the updates of the estimates are implemented. In (Samek et al., 2014) the authors opted for applying multiplicative updates that require the determination of the gradient of the criterion in the space of skew-symmetric matrices, whereas our proposal performs tangent updates to the manifold of the semi-orthogonal matrices that are followed by a projection or retraction onto the manifold. These updates are quite common in the research field of Independent Component Analysis (Edelman et al., 1998; Amari, 1998; Cruces-Alvarez et al., 2004; Nishimori, 1999).

The main steps of the Sub-ABLD iteration are summarized in Algorithm 2.

Algorithm 2 Sub-ABLD algorithm

function SUB-ABLD($\{\mathbf{P}_j\}$,$\{\mathbf{Q}_j\}$,p,α,β,η)

Compute the average covariance matrices $\mathbf{P}$ and $\mathbf{Q}$.

Compute the total covariance matrix $Cov(\boldsymbol{x}) = p(c_1)\mathbf{P} + p(c_2)\mathbf{Q}$.

Compute the whitening transform matrix $\mathbf{T}$ using Eqn. 7.8.

Whiten the trial and average covariance matrices to respectively obtain $\{\breve{\mathbf{P}}_j\}$,$\{\breve{\mathbf{Q}}_j\}$ and $\breve{\mathbf{P}}$,$\breve{\mathbf{Q}}$.

Compute the scaling parameter, κ using Eqn. 6.62 and initialize the iteration counter: $i = 0$.

Initialize the semi-orthogonal matrix $\mathbf{\Omega}^{(i)} = \mathbf{I}_{n\times p}$.

repeat

Compute the robust criterion $f(\mathbf{\Omega}^{(i)})$ using Eqn. 7.11).

Compute the ordinary gradient $\frac{\partial f(\mathbf{\Omega}^{(i)})}{\partial \mathbf{\Omega}}$ using Eqn. 7.12.

Compute the natural gradient on the Stiefel manifold $\nabla_\Omega f(\mathbf{\Omega}^{(i)})$ using Eqn. 7.13.

Obtain the tangent matrix $\mathbf{\Omega}_{tg}^{(i+1)}$ using Eqn. 7.14.

Obtain the projection matrix $\mathbf{\Omega}^{(i+1)}$ using Eqn. 7.16 and Eqn. 7.17 (the retraction onto the manifold).

Increase the iteration counter: $i = i + 1$.

until convergence at iteration i_{max}.

Collect in $\breve{\mathbf{V}}$ the princip./minor eigenvect. of the pencil $((\mathbf{\Omega}^{(i_{max})})^T\breve{\mathbf{P}}\mathbf{\Omega}^{(i_{max})}, (\mathbf{\Omega}^{(i_{max})})^T\breve{\mathbf{Q}}\mathbf{\Omega}^{(i_{max})})$.

return $\mathbf{W}^T = \breve{\mathbf{V}}^T(\mathbf{\Omega}^{(i_{max})})^T\mathbf{T}$.

end function

7.3 Experimental Study

The discrimination of two class MI movements consists of the following steps. The MI EEG signals are acquired, preprocessed and spatially filtered. These filtered signals are then used for extracting the required features, which are classified using a linear classifier. In the following section, the experimental steps used for testing is explained and the proposed algorithm is compared with standard CSP.

7.3.1 Simulations Data and Preprocessing

Initially, the robustness of the proposed algorithm is explored in a controlled situation with synthetic data. Two sets of SPD that represent the trial covariance matrices of the two classes were randomly generated. Each set consists of 200 trials. For further preprocessing, both the sets of matrices were concatenated. The concatenated data are cross-validated using k-fold Cross-Validation (CV) $(k = 10)$. This divides the data into 10 equal subsets in which a single set was used as a testing data and the remaining 9 subsets were used for training the classifier. The performance of the proposed algorithm was studied in the presence of the outliers. The outliers consist of matrices with abnormal higher variances that were inserted in the training set of both the classes. The proposed Sub-ABLD algorithm was tested in Fig. 8.7 by progressively varying the percentage of outliers in the trials from 0% until 30%. The robustness of Sub-ABLD and its comparison with respect to the other algorithms mentioned in the figure will be addressed in Section 7.4.

7.3.2 EEG Dataset and Preprocessing

To evaluate the proposed Sub-ABLD algorithm with BCI competition datasets, two datasets from competition III: dataset 3a, dataset 4a (which can be downloaded from (*BCI Competition III*, 2005)) and one dataset from competition IV data set 2a (which can be downloaded from (*BCI Competition IV*, 2008)) were utilized. The data were acquired during the MI movements. The first dataset 3a (Schlögl et al., 2005) from BCI competition III (Blankertz et al., 2006), were acquired from 3 healthy subjects namely K3, K6 and L1 using 60 channels EEG acquisition system. The signals were recorded while executing the MI movements of the left hand, right hand, foot and tongue. The signals were sampled at a frequency of 250 Hz. The sampled signals were bandpass filtered at the frequency range between 1 to 50 Hz. The data set consists of two sessions i.e., training and testing sessions. For subject K3, both the sessions consist of 45 trials for each class whereas the other two subjects i.e., K6 and L1 performed 30 trials per class in both the sessions. For the second dataset, data set 4a (Dornhege et al., 2004) of BCI competition III (Blankertz et al., 2006), the signals were acquired from five subjects namely AA, AL, AV, AW and AY using 118 channels EEG system. The acquisition was done during the imagery movements of the left hand, right hand and right foot. Down-sampling of the recorded signals was done at 100 Hz. The band-pass filter between 0.05 to 200 Hz frequency band was applied to the signals. The data set of each subject consists of 280 total trials. The size of the training sessions is different from testing sessions. The training sessions consist of 168, 224, 84, 56, 28 trails for subjects AA, AL, AV, AW, AY and the remaining denotes the testing trails for the corresponding subjects. The last dataset, data set 2a (Naeem et al., 2006) BCI competition IV (Tangermann et al.,

2012) were acquired from nine subjects (A1 to A9) while performing the left hand, right hand, foot and tongue MI movements using 22 electrodes. The sampling frequency of the signals was 250 Hz. The band-pass filtering of the acquired signals was performed between 0.5 and 100 Hz. For each subject, the data were acquired on different days and each set consists of 72 trials for each class.

In this approach, the performances were obtained using only two MI movements considering all the channels from each dataset. The preprocessing step was implemented similarly for all the algorithms. First, a fifth-order band-pass filter with a cut-off frequency between 8 to 30 Hz was applied to the raw EEG signals. A time window of 2s during the imagination of movements was extracted for each trial. The extracted trials were concatenated for each class and applied a k-fold cross-validation $(k = 10)$ to the concatenated data. The CV process divides the data into 10 equal sets where one set of data was used as testing data and the remaining 9 sets were used for training. Finally, the optimal spatial filters were obtained using the training dataset. The number of filters selected for each class is $k = 3$, so the total number $p = 6$.

7.3.3 Feature Extraction and Feature Classification

For both-the synthetic and the BCI datasets, the obtained spatial filters were used for filtering the training and testing data. The training and testing features were obtained by taking the log-variance of the filtered data in order that their distribution be closer to Gaussianity. The LDA (Duda and Hart, 1973) classifier was used for discriminating the features of the two classes. The classifier was trained using the training features and its performance was obtained using the testing features. The preprocessing, feature extraction and classification steps were repeated 10 times and finally the average performance was obtained.

7.3.4 Selection of α, β and η Values

The selection of α and β is one of the crucial steps for the proposed algorithm. Depending on the α and β values, the AB Log-Det divergence can be derived into different divergence techniques (Cichocki et al., 2015). The proposed algorithm performed better when $\alpha = \beta$, a situation where the AB Log-Det divergence is symmetric or invariant under the permutation of its arguments. In this experiment, the performance for various values of $\alpha = \beta$ and η was observed, and a suitable configuration of parameters for each dataset was selected.

7.4 Results and Discussion

The performance of the proposed Sub-ABLD algorithm is compared with the performance of the CSP algorithm for both the synthetic and the real BCI competition datasets.

In order to carry out a fair performance comparison, a total of six features (i.e., $p = 6$) have been selected for both the algorithms. The performance comparison between both the algorithms is presented in the following subsections.

7.4.1 Observations for Simulated Data

To study the performance of the proposed algorithm in the presence of outliers, the experiment was done by increasing the percentage of outlier trials in the training set for both the classes.

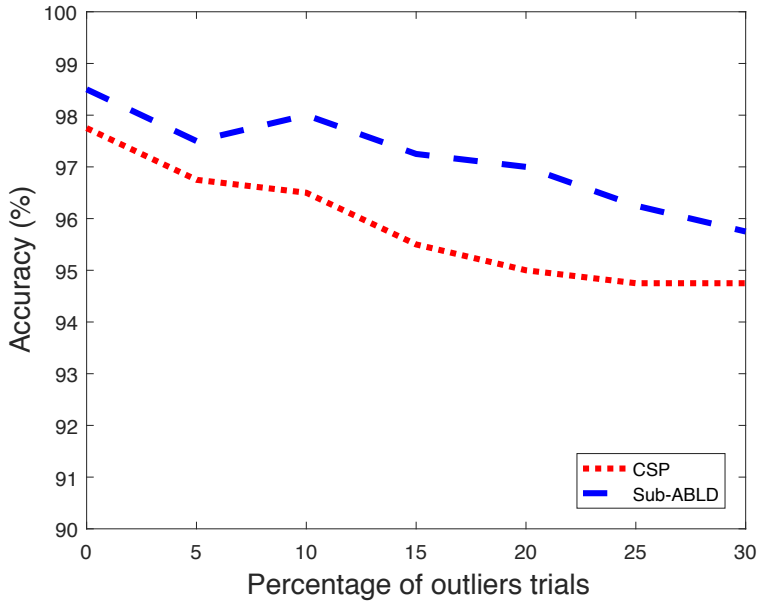
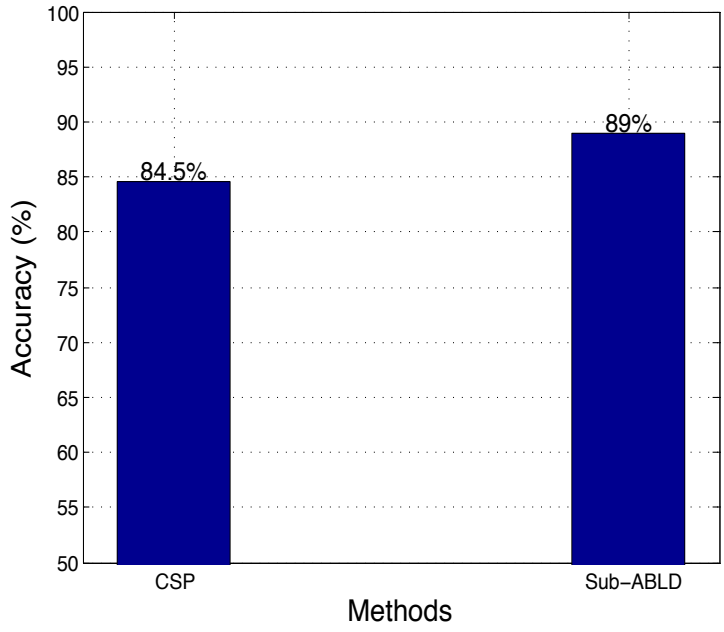


Fig. 7.1 Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 1,\ \alpha = \beta = 1.5$) with CSP versus the percentage of outlier trial

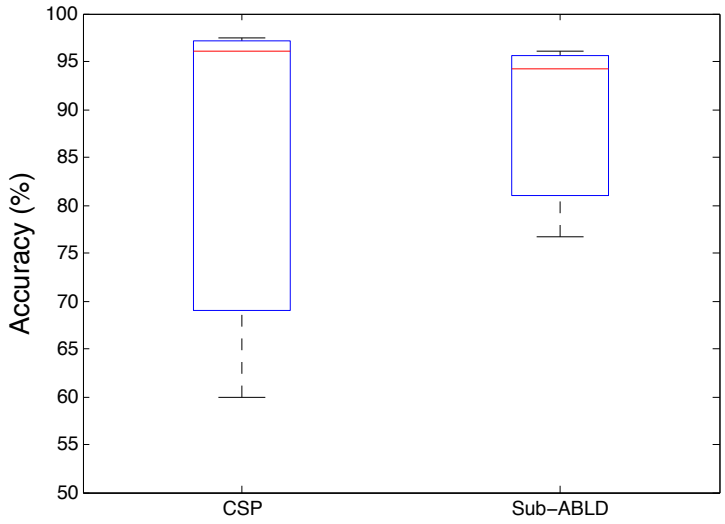
The performances of the above algorithms with the increasing percentage of outliers in the training set are presented in Fig. 7.1. It can be observed that CSP performs worse in the presence of the outliers than the proposed Sub-ABLD algorithm.

7.4.2 Observations for BCI Competition Datasets

In this section, the proposed algorithm is tested using three BCI competition datasets. For each dataset, the performances of the proposed algorithm for the different values of (α, β) and η were observed. From the observation, the maximum performance of the Sub-ABLD algorithm for the particular (α, β) and η values was selected. The selected performance is compared with the performances of other existing algorithms. Further analysis is done by using a box plot comparison for all the algorithms. The box plot analysis shows the distribution of the performances. In a box plot representation, the line inside the box represents the median performance. The upper and lower hinge of the

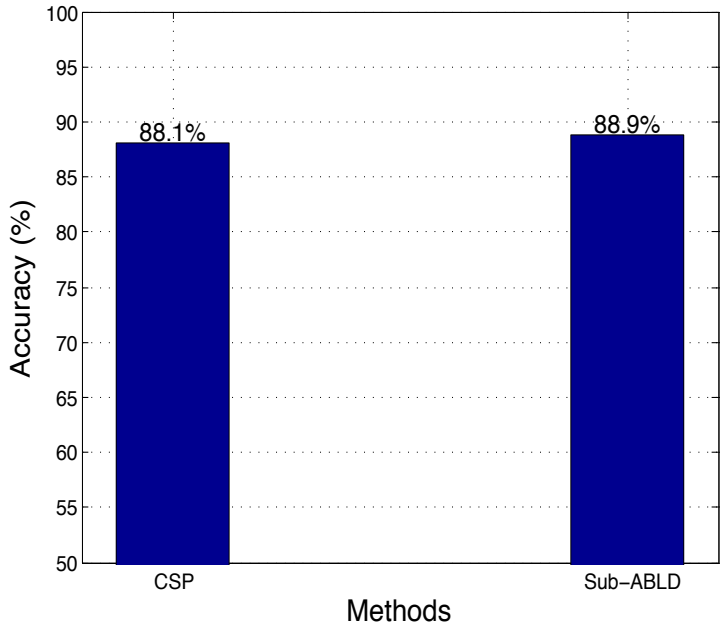


(a)

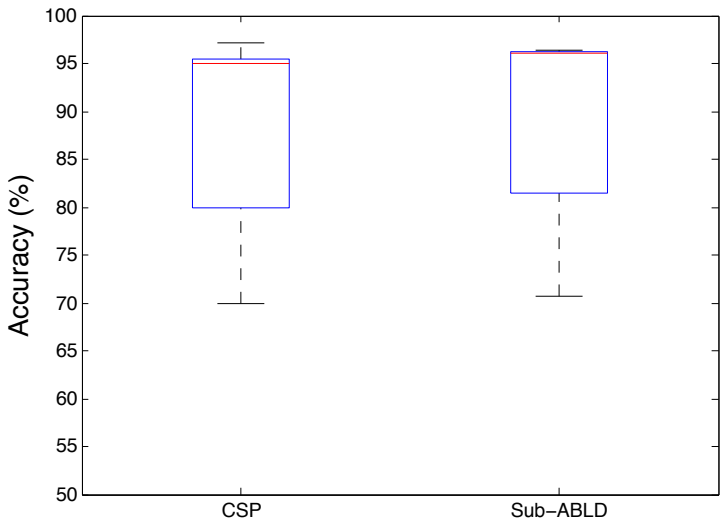


(b)

Fig. 7.2 (a) Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 2,\ \alpha = \beta = 1.5$) with CSP using BCI competition III dataset 3a and (b) its corresponding boxplot.

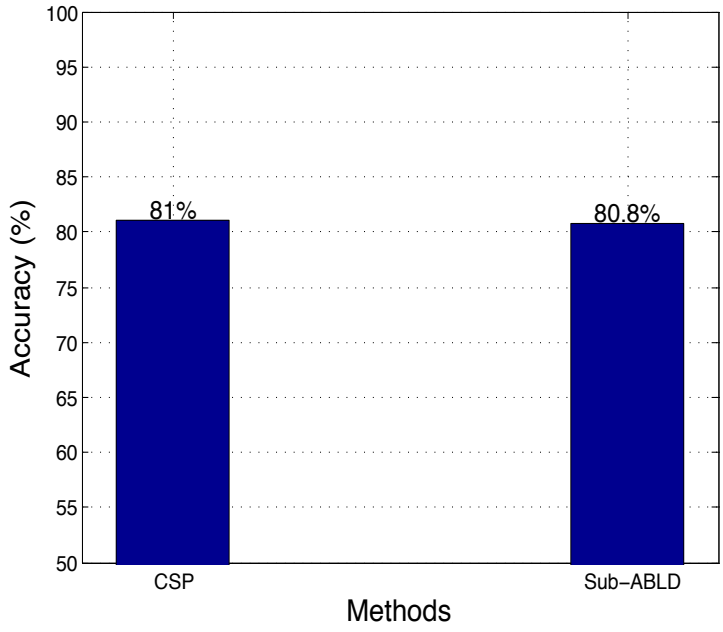


(a)

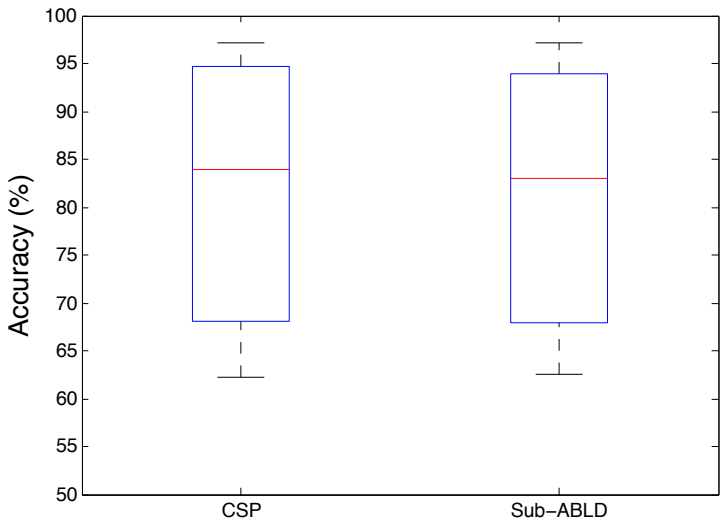


(b)

Fig. 7.3 (a) Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 0.5,\ \alpha = \beta = 2$) with CSP using BCI competition datasets III dataset 4a and (b) its corresponding boxplot.



(a)



(b)

Fig. 7.4 (a) Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 0.25,\ \alpha = \beta = 1.25$) with CSP using BCI competition datasets IV dataset 2a and (b) its corresponding boxplot.

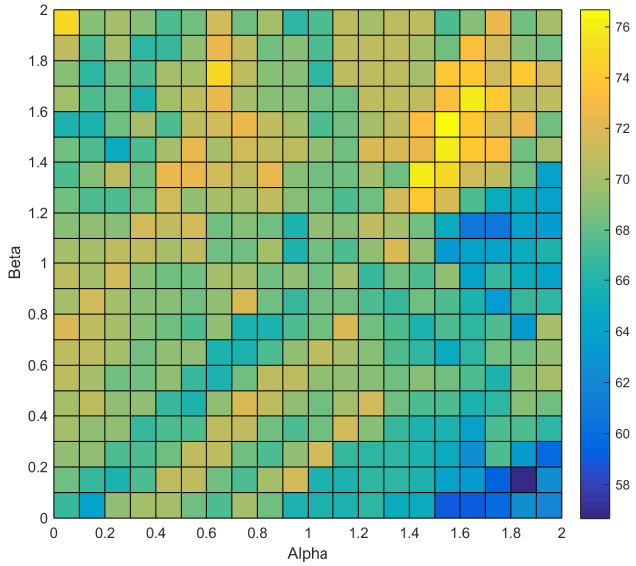


Fig. 7.5 Results of the Sub-ABLD algorithm for the subject k6 from BCI competition III dataset 3a. This figure illustrates the changes in the average classification performance with respect to the variation of the parameters α and β. Relatively good performance results are obtained close to the diagonal and for moderately large values of the parameters.

box denote the 75^{th} and 25^{th} percentile of the overall performance distributions. The whiskers are symbolized by the two lines outside the box. The upper and lower whisker represents the maximum and minimum performance observed.

For BCI competition III dataset 3a, the Fig. 7.2a shows the comparison of the highest average performance of the Sub-ABLD algorithm with the average performances of other existing algorithms. From the figure, it is observed that the Sub-ABLD algorithm outperforms the CSP algorithm with an average performance accuracy of 89% for this dataset. The box plot comparison is shown in Fig. 7.2b. Although the median performance is slightly higher for CSP, the 25^{th} percentile performance is much smaller than the one of the Sub-ABLD algorithm. As it will be seen later, is a consequence that with the Sub-ABLD algorithm the most difficult subjects have attained a significant improvement in their classification performance.

Fig. 7.3 shows the observed average performances using BCI competition III dataset 4a. For this dataset, the proposed algorithm Sub-ABLD performs slightly above than the average performance of CSP, which is 88.1%. From the box plot of the results, it can be observed that the 25^{th} percentiles for both the algorithms are also quite close.

Similar results have been obtained for the BCI competition IV dataset 2a, which is shown in Fig. 7.4. Again the algorithms performance of Sub-ABLD is same as CSP, which average performance is 81%. In the box plot, we can observe that the quartiles of both the algorithms coincide.

To analyze the effect of performance for different divergences, the parameters (α, β) is varied for a single subject (Subject k6 from BCI competition III dataset 3a, which is one of the subjects with a worst performance for the experiment) and obtained the corresponding performance. The values of (α, β) are varied to cover the interval $[0, 2] \times [0, 2]$ with a mesh of 0.1 spacings. The observed performance is shown in Fig. 7.5. This figure reveals a tendency to improve the classification accuracy of the worst user for values of α and β that are close to the diagonal and large enough so they can effectively down-weight the contribution in the estimating equations coming from the largest and smallest generalized eigenvalues.

7.5 Conclusions

The Sub-ABLD algorithm has been proposed by optimizing the proposed criterion based on AB Log-Det divergence for discrimination of two class imagery movements. The spatial filters are computed by considering both. This algorithm was tested with synthetic and real datasets and compared with the standard CSP algorithm. The simulations have confirmed the possibility to tune up the hyperparameters of the divergence so as to improve the robustness of the obtained solutions without deteriorating the expected accuracy.

CHAPTER 8

Simulations

This thesis presents two algorithms: ThinICA-CSP algorithm which aims at discriminating four class MI movements and Sub-ABLD for discriminating two class MI movements. The comparison of the performances of the proposed algorithms with the baseline method was already presented in chapter 5 and chapter 7. In this chapter, the performance of the proposed algorithms is compared with existing algorithms. All the coding and execution are performed by using MATLAB 2013a version. Section 8.1 presents the simulation study and comparison of the performance obtained using ThinICA-CSP algorithm. An experimental study of Sub-ABLD is discussed in Section 8.2.

8.1 Simulations using ThinICA-CSP algorithm for discrimination of four class motor imagery movements

The comparison of the performance of ThinICA-CSP with the other algorithms like mCSP and JADE for discrimination of four class movements are presented in this section. mCSP is an extension of the standard CSP by regarding the multiclass problem as binary problems for computation of spatial filters (Dornhege et al., 2004). JADE algorithm performs a joint approximate diagonalization of the trial covariance matrices of the classes (Grosse-Wentrup and Buss, 2008).

8.1.1 Experimental Set-up

The dataset that was used in Chapter 5 is being considered here. The experimental study is also similar to that explained in section 5.3. For this study, the JADE algorithm is included for a comparative study of performance. The filtered signals are used for the computation of spatial filters using the multiclass CSP, JADE method and the proposed ThinICA-CSP algorithm. Similar to the process presented in chapter 5, two spatial filters have been selected for each class which gives a total of $p = 8$ spatial filters for four classes. The training and testing signals were filtered using the obtained spatial filters. The log transformation of the variance of the spatially filtered signals gives the required features. The extracted training features were used for training LDA classifier. The overall experimental study is shown in Fig 8.1

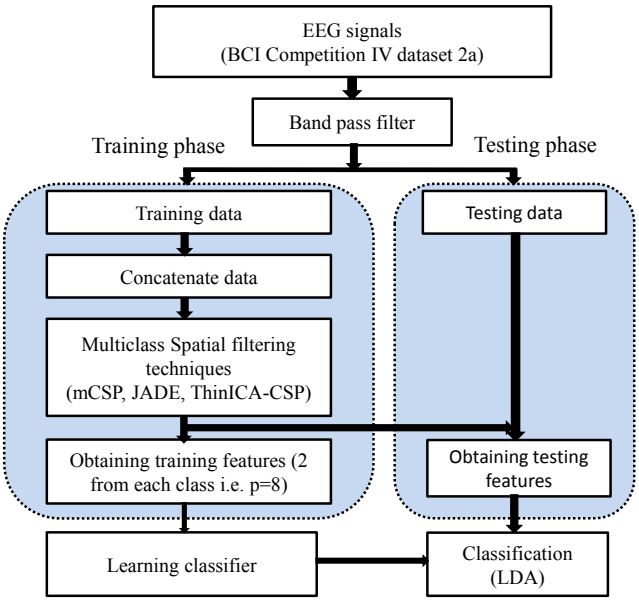


Fig. 8.1 Experimental study for performance comparison of mCSP, JADE and ThinICA-CSP

8.1.2 Performance Results

The performance results of ThinICA-CSP based on the above experimental set up are presented here. Fig. 8.3 shows one spatial pattern of the selected features during MI movements of the left hand, right hand, foot and tongue for subject A1 . The first, second and third column in Fig. 8.3 represents, respectively the spatial patterns obtained using ThinICA-CSP, JADE and multiclass CSP. In the first column of the figure, it can be observed that the ERD activity occurs in the right motor cortex during the left hand MI movement. Similarly, the right hand MI movement results in the ERD activity in the left motor cortex, foot motor imagery shows both ERS and ERD activities in the mid-central region and tongue MI movement denotes more ERS activity on the parietal region. Thus, the above observations agree with the findings in (Pfurtscheller et al., 2006). But, the patterns obtained using JADE and multiclass CSP are not much in agreement with the above findings. This shows that the ThinICA-CSP selects more relevant features than the above mentioned methods. For further analysis, the performance accuracies of the above methods are compared using the same dataset. The performance results are shown in Fig. 8.4. From the figure, it is observed that the ThinICA-CSP gives the highest performance of 64% than the multiclass CSP and JADE algorithms.

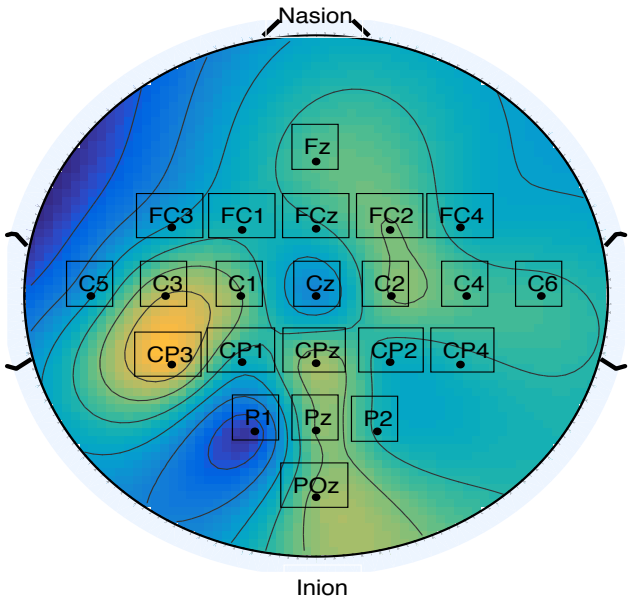


Fig. 8.2 Legends of the electrode locations used for acquisition of EEG from BCI competition IV dataset 2a along with nasion and inion.

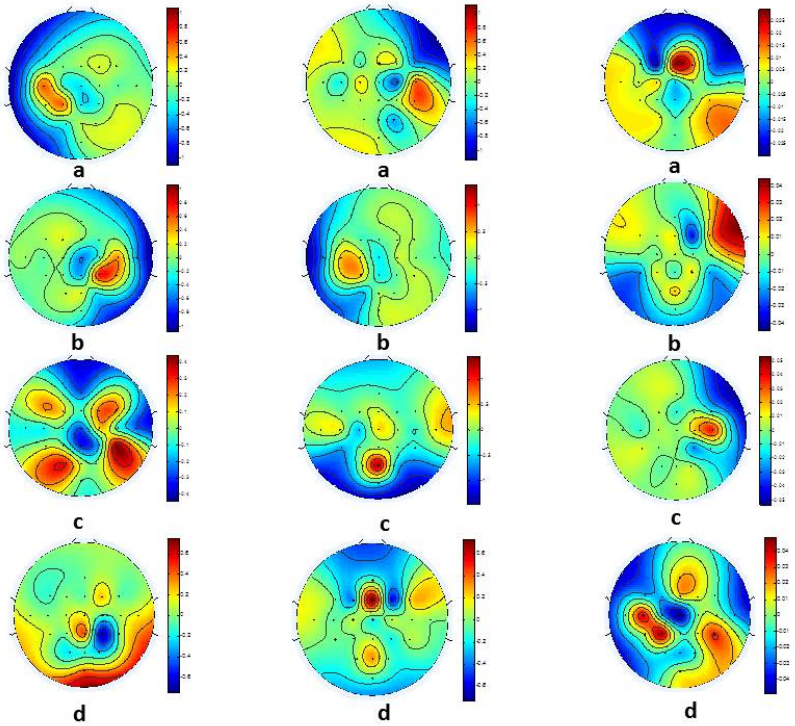


Fig. 8.3 Spatial pattern obtained during MI movements of (a) left hand, (b) right hand, (c) foot and (d) tongue for subject A1 from BCI competition IV dataset 2a using ThinICA-CSP, JADE and multiclass CSP

The box plot comparison of all the three methods is shown in Fig. 8.5. The median performance is represented by the middle line inside the box. The 75^{th} percentile and

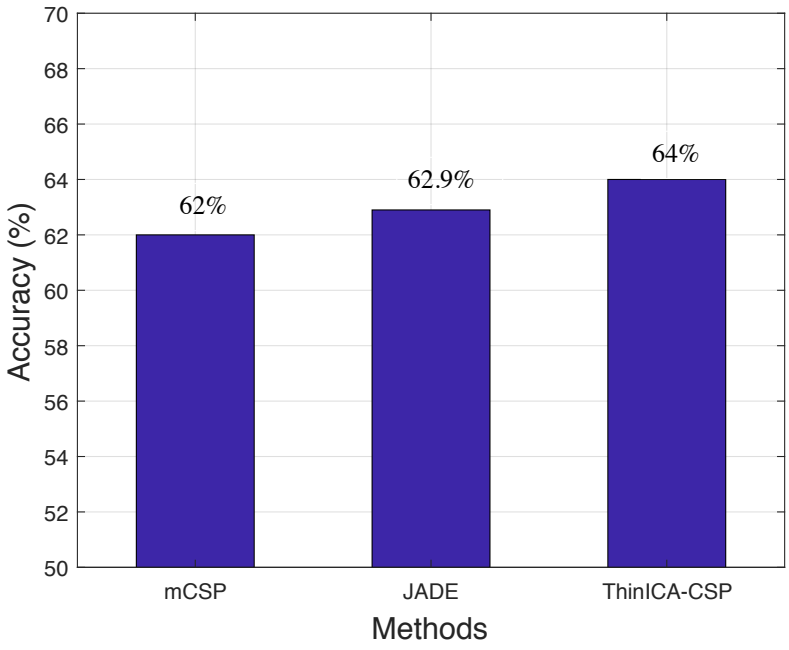


Fig. 8.4 Comparison of classification performance for mCSP, JADE and ThinICA-CSP

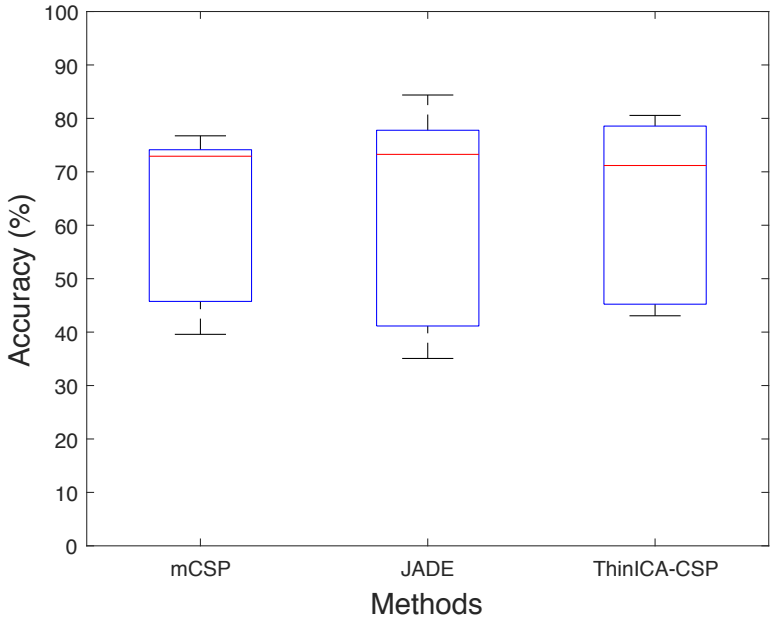


Fig. 8.5 Boxplot comparison of mCSP, JADE and ThinICA-CSP

25^{th} percentile of the overall performance of each method is denoted by the upper edge and lower edge of the box respectively. The outliers are indicated by whiskers. The median is similar for all the compared methods but the 25^{th} percentile is little higher for ThinICA-CSP as compared to multiclass CSP and JADE. Thus, this indicates that the ThinICA-CSP performs better for a subject whose performance is worse with the other two algorithms. Moreover, the subject who performs better with the other two methods also performs better using ThinICA-CSP which is shown by the increase in 75^{th}.

8.2 Simulations using Sub-ABLD algorithm for discrimination of two class MI movements

In this section, the proposed Sub-ABLD algorithm has been studied in various scenarios. Firstly, the robustness of the proposed algorithm is tested and compared with various existing algorithms. Secondly, different experiments were done such as computation of the mean with two different methods, including the oultier in BCI competition datasets, imbalancing the number of training trials for both the classes and varying the number of training sets used for training the classifier.

8.2.1 To Study the Robustness of the Proposed Algorithm and Compare its Performance with the Other Existing Algorithms

The performance obtained using Sub-ABLD is compared with other existing algorithms such as standard CSP, JADE, MAPCSP and divCSP-WS. JADE algorithm performs a joint approximate diagonalization of the trial covariance matrices of the classes (Grosse-Wentrup and Buss, 2008). MAPCSP is a Bayesian algorithm that tries to find the maximum a posteriori estimates of the patterns and sources in a generative model with additive Gaussian isotropic noise (Wu et al., 2015). The subspace implementation of divCSP-WS finds a balance between the maximization of Beta divergence between the conditional covariances of the filtered outputs for each class and the minimization of the variability within each class (Samek et al., 2014). This algorithm contains two hyper-parameters, the regularization factor λ and the real scalar β' that specifies the chosen Beta divergence. The factor λ admits an equivalence in terms of the regularization parameter η in Sub-ABLD which link them through the mapping $\lambda \equiv \eta/(1+\eta)$, while the parameter of the Beta divergence β'_* was chosen in the simulations to maximize the performance. We have performed two experiments. The first experiment is to study the robustness of the proposed algorithm in the presence of the outlier trials using artificial data. The second experiment is to compare the performance of the above mentioned algorithm with three BCI competition datasets.

8.2.1.1 Experimental set up using Simulated and Real Dataset

This experiment is done on both the synthetic and the real BCI competition datasets. To study the robustness of the proposed algorithm in the presence of outliers, the artificial data generated in section 7.3.1 was used. The experiment steps are performed similarly as described in section 7.3. For real dataset, we have used three BCI competition datasets described in section 7.3.2 and the experimental study is done similarly as described in chapter 7. In order to carry out a fair performance comparison, a total of six features (i.e., $p = 6$) have been selected for all the algorithms. The implementation of the JADE and divCSP-WS algorithms were taken from the web pages of the authors. The baseline divCSP-WS algorithm has been downloaded from (*The Divergence Methods Web Site*, 2013), while the implementation of JADE algorithm can be found at (*Machine Learning in Neural Engineering*, 2017(last update)). The experiment is done using all the above mentioned algorithms. The overall experimental study is shown in Fig 8.6

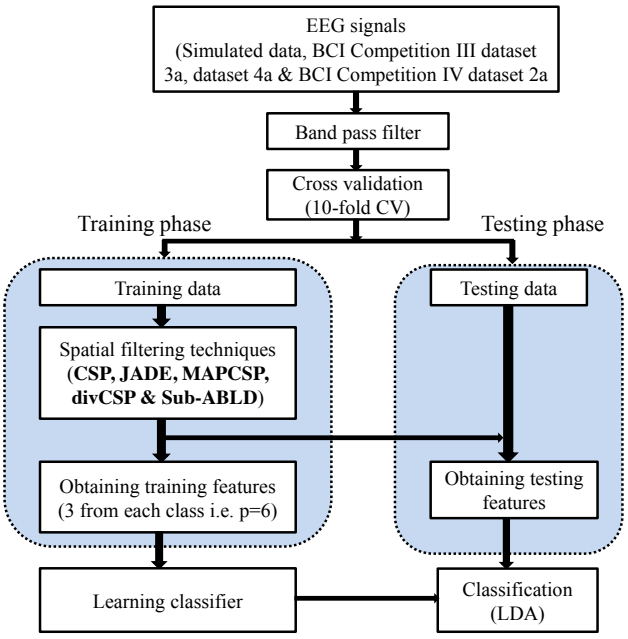


Fig. 8.6 Experimental study for performance comparison of CSP, JADE, MAPCSP, divCSP and Sub-ABLD

8.2.1.2 Performance Results for Simulated Data

The performances of the above algorithms with the increasing percentage of outliers in the training set are presented in Fig. 8.7. It can be observed that MAPCSP performs worse in the presence of the outliers. The performances of CSP, JADE and divCSP-WS are much more robust than MAPCSP, but in overall, the proposed Sub-ABLD algorithm seems to outperform the compared algorithms in the presence of the outliers.

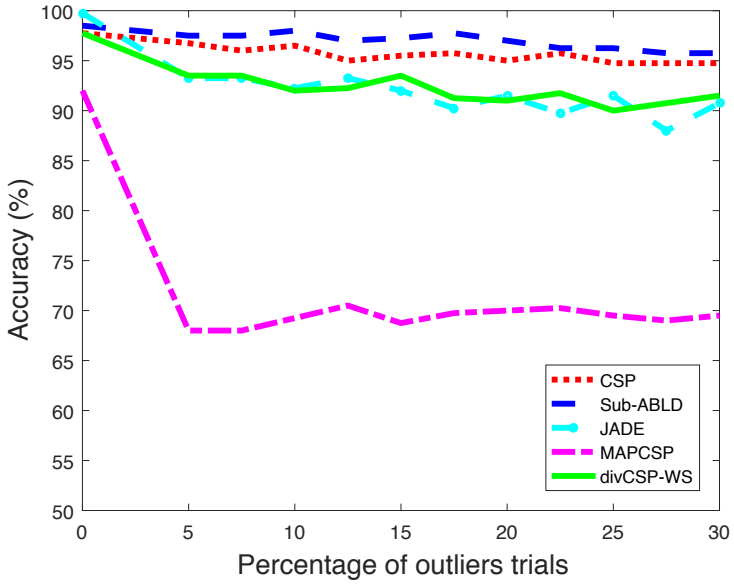


Fig. 8.7 Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 1,\ \alpha = \beta = 1.5$) with CSP, JADE, MAPCSP and divCSP-WS ($\lambda = 0.5, \beta'_* = 0.25$), versus the percentage of outlier trial

8.2.1.3 Performance Results for BCI Competition Datasets

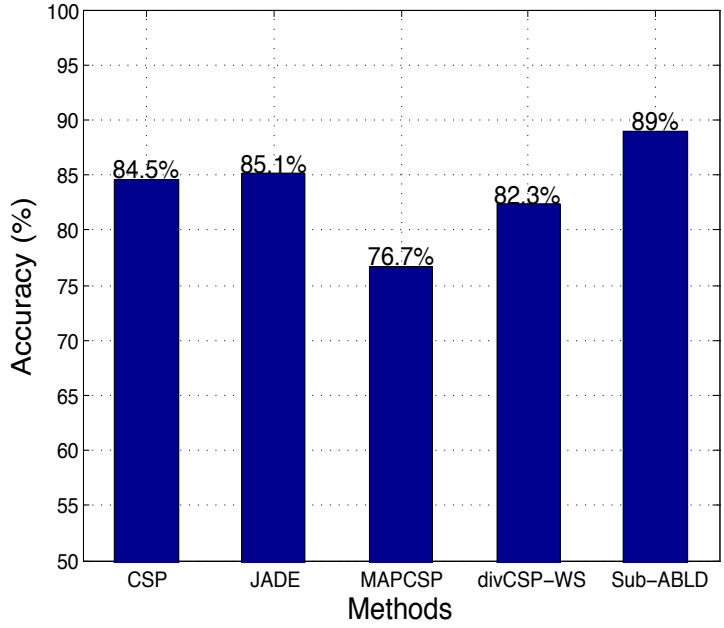
For BCI competition III dataset 3a, Fig. 8.8a shows the comparison of the highest average performance of the Sub-ABLD algorithm with the average performances of other existing algorithms. From the figure, it is observed that Sub-ABLD algorithm outperforms the other existing algorithms with an average performance accuracy of 89% for this dataset. The box plot comparison is shown in Fig. 8.8b. Although, the median performance is slightly higher for CSP, JADE and divCSP, their 25th percentile performance is much smaller than the one of the Sub-ABLD algorithm. As we will see later, is a consequence that with the Sub-ABLD algorithm the most difficult subjects have attained a significant improvement in their classification performance.

Fig. 8.9 shows the observed average performances using BCI competition III dataset 4a. For this dataset, the algorithms JADE, Sub-ABLD and divCSP-WS perform essentially similar and slightly above than the average performance of CSP, which is 88.1%. From the box plot of the results, we can observe that the 25th percentiles for these four algorithms are also quite close.

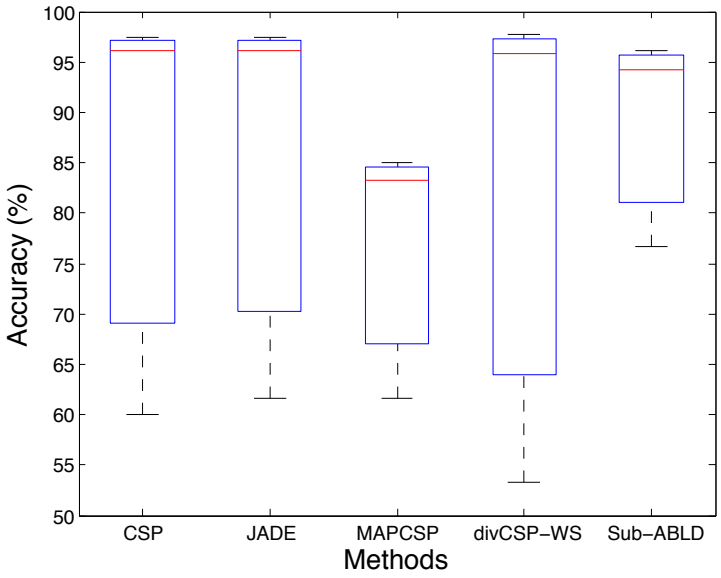
Similar results have been obtained for the BCI competition IV dataset 2a, which is shown in Fig. 8.10. Again the algorithms JADE, Sub-ABLD and divCSP-WS perform essentially the same as CSP, which average performance is 81%. In the box plot, we can observe that the quartiles of these algorithms are approximately coincident.

The proposed Sub-ABLD algorithm has been tested on both simulated and real EEG signals. On one hand, the results with synthetic data indicate that the proposed Sub-ABLD exhibits a certain robustness to the presence of outlier trials in the dataset.

On the other hand, the analysis of real EEG signals is also challenging because of the possible presence of artifacts and non-stationarities. We have presented the performance of the Sub-ABLD algorithm using several real BCI datasets. For BCI competition III dataset 3a, we can observe that the proposed Sub-ABLD algorithm also outperforms the other algorithms. Whereas, the performance of the proposed algorithm is almost similar to the one obtained by JADE, divCSP-WS and CSP in the other two datasets, i.e., for the BCI competition III dataset 4a and BCI competition IV dataset 2a. Additionally, the analysis of the box-plots reveals that the proposed Sub-ABLD algorithm increased the classification performance of the subjects that do not perform well for the other methods. At the same time, it retained an almost similar performance for the remaining subjects. These observations meet our initial goal of developing a robust algorithm. The classification performance is also affected by the regularization parameter η that controls the penalty term. In general, the data with outliers give the best performance for the higher values of η and, otherwise, smaller values are preferable. In this study, the value of η has been kept constant across subjects in each dataset.

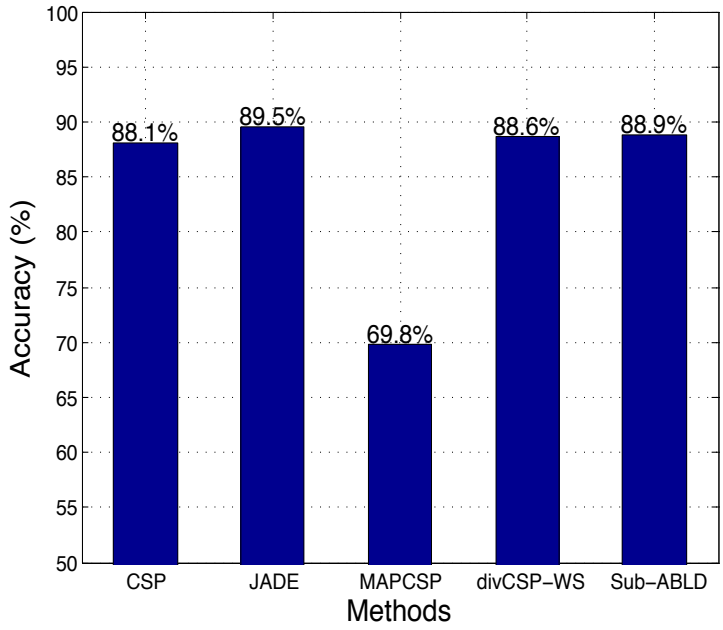


(a)

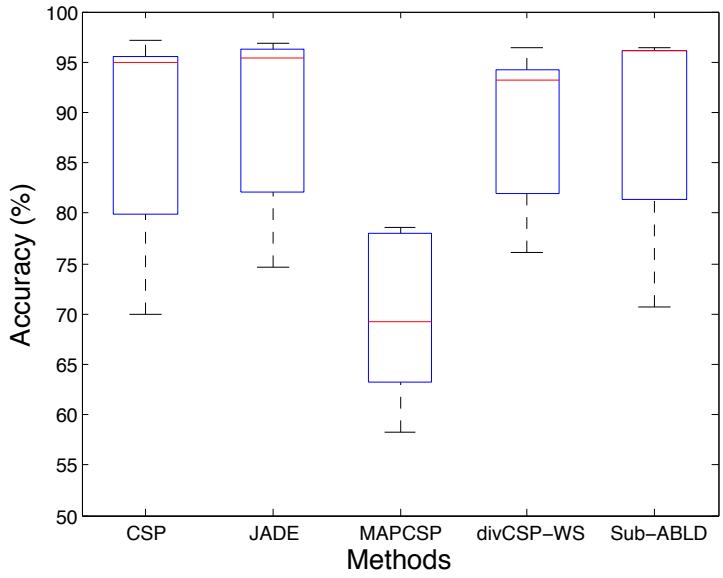


(b)

Fig. 8.8 (a) Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 2,\ \alpha = \beta = 1.5$) with CSP, JADE, MAPCSP and divCSP-WS ($\lambda = 0.66,\ \beta'_* = 1$) using BCI competition III dataset 3a and (b) its corresponding boxplot.

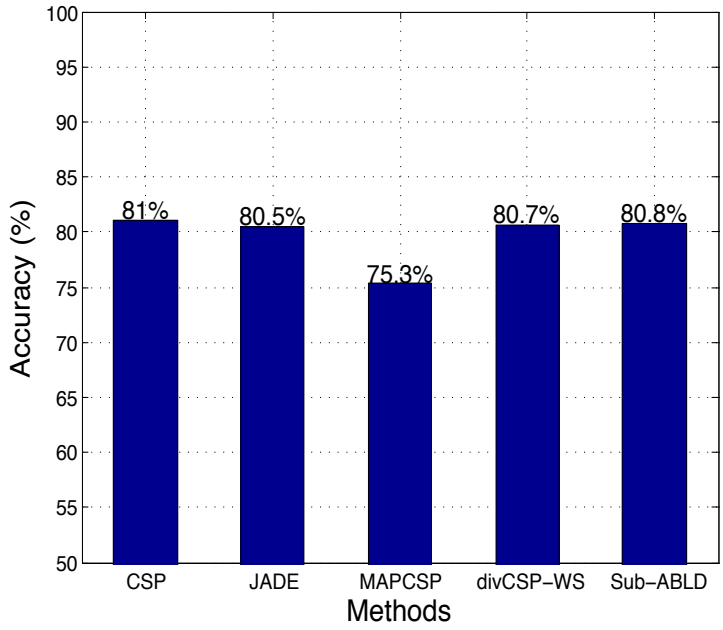


(a)

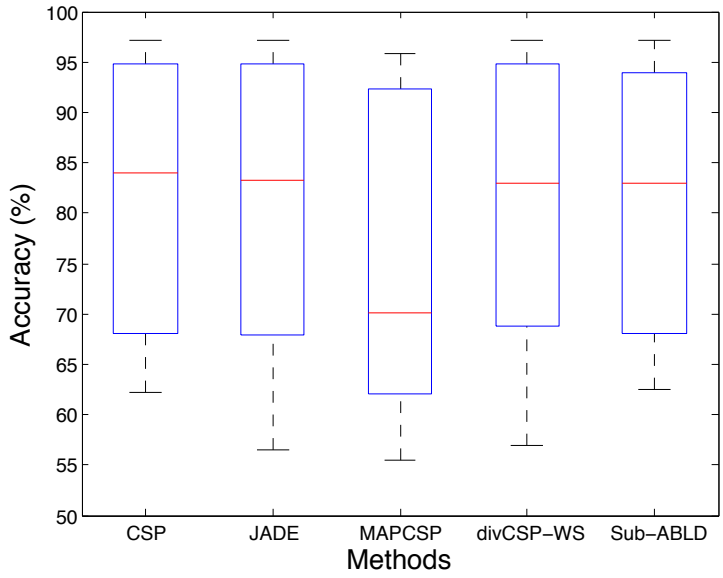


(b)

Fig. 8.9 (a) Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 0.5,\ \alpha = \beta = 2$) with CSP, JADE, MAPCSP and divCSP-WS ($\lambda = 0.33,\ \beta'_* = 0.5$) using BCI competition datasets III dataset 4a and (b) its corresponding boxplot.



(a)



(b)

Fig. 8.10 (a) Performance comparison of the proposed algorithm Sub-ABLD ($\eta = 0.25,\ \alpha = \beta = 1.25$) with CSP, JADE, MAPCSP and divCSP-WS ($\lambda = 0.2,\ \beta'_* = 0$) using BCI competition datasets IV dataset 2a and (b) its corresponding boxplot.

8.2.2 To Study the Performance of the Proposed Algorithm in Different Scenarios

The following experiments are performed on CSP and the proposed algorithm.

(1) Computation of average using Arithmetic and Geometric mean: Arithmetic mean is commonly used for averaging the distributions. However, considering the covariance

manifold which is not exactly a plane surface, averaging using geometric mean seems to be more meaningful and accurate.

(2) Including outliers in the training trials: Oultier plays a major role in the computation of the efficiency of a system. Therefore, to study the performance of the algorithms, artificial outliers with large variance are included in the EEG dataset. The outlier matrices are generated randomly by drawing a Gaussian matrice $\mathbf{A^{out}}$ with i.i.d. elements and obtained $Cov_{out} = A^{out}(A^{out})^T$. Later, the generated Cov_{out} is scaled by a factor of 1.

(3) Imbalancing the number of training trials for two classes: Generally, equal number of training trials for both the classes are considered for the computation of spatial filters. Therefore, to observe the changes in the performance by imbalancing the number of the training trials, unequal number of training trials for class 2 is considered.

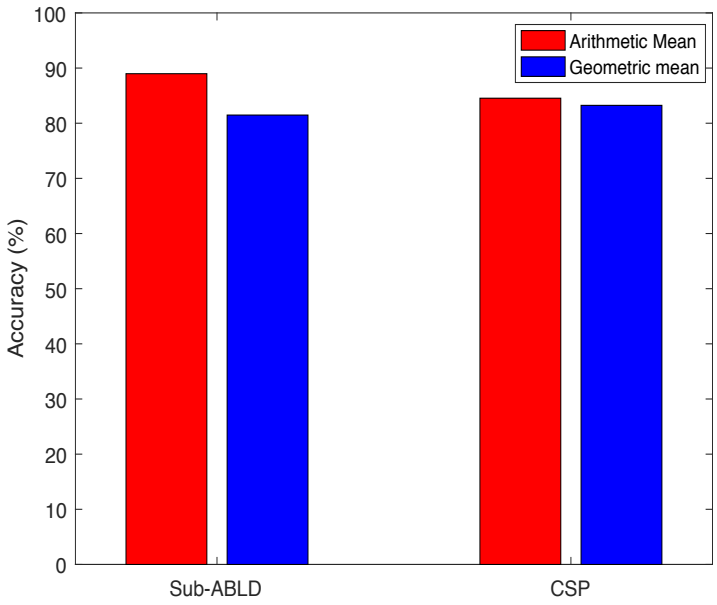
(4) Varying the number of training trials: The number of trials used for training the system influences both the performance as well as the processing time. Hence, to select the minimum numbers of trials without compromising the accuracy and the processing time, the experiment is performed by varying the k-folds value.

8.2.2.1 Experimental Set-up

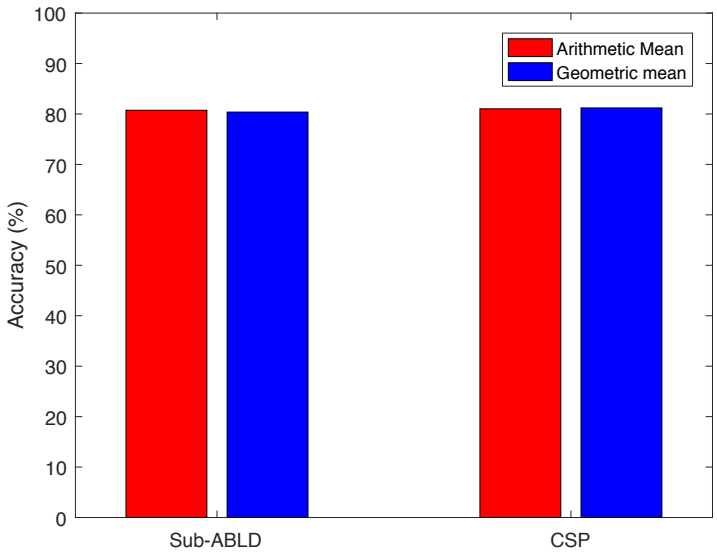
These experiments are done using BCI competition III dataset 3a and competition IV dataset 2a.The preprocessing of the data is done similar to the steps described in Section 7.3.2. In the first experiment, both the Arithmetic and Geometric mean are used for computing the average covariance matrices and the performance obtained is compared. The second experiment is done by including the artificial outlier trials in BCI competition dataset and observing the changes in performance by increasing the number of outlier trials. The third experiment is done by varying the number of trials of the two classes used for the computation of spatial filters. The last experiment is done by varying the number of training trials. This is done by varying the k values from [k=5,. . . ,30].

8.2.2.2 Performance Results

All the performance results obtained using BCI competition III dataset 3a and BCI competition IV dataset 2a are presented. The performance results with different methods for averaging the class covariance matrices are shown in Fig. 8.11. From the figure, it is observed that for BCI competition III dataset 3a, the Arithmetic mean performs better than the Geometric mean, whereas both the methods perform almost similar for BCI competition IV dataset 2a. Fig. 8.12 represents the performance of the algorithms in the presence of the outliers. For this experiment, the performance is observed by increasing the number of outlier trials in both the class covariance matrices. The performance degrades with the increased in the number of outlier trials for both the algorithms.

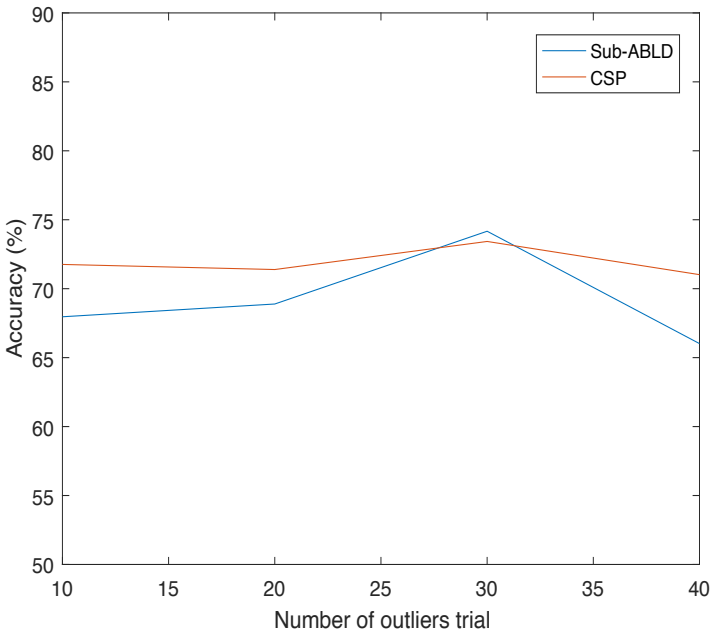


(a)

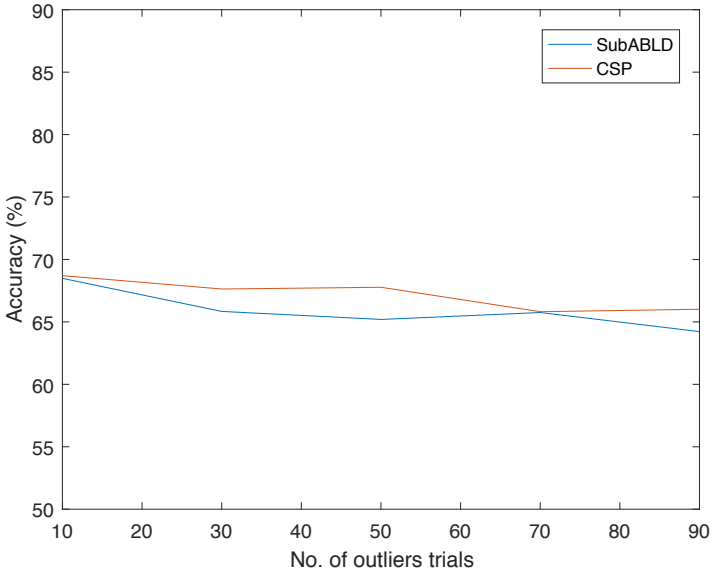


(b)

Fig. 8.11 (a)Performance comparison of Sub-ABLD algorithm ($\eta = 2,\ \alpha = \beta = 1.5$) and CSP with Arithmetic mean and Geometric mean using BCI competition datasets III dataset 3a and (b) Performance comparison of Sub-ABLD algorithm ($\eta = 0.25,\ \alpha = \beta = 1.25$) and CSP with Arithmetic mean and Geometric mean using BCI competition datasets IV dataset 2a.

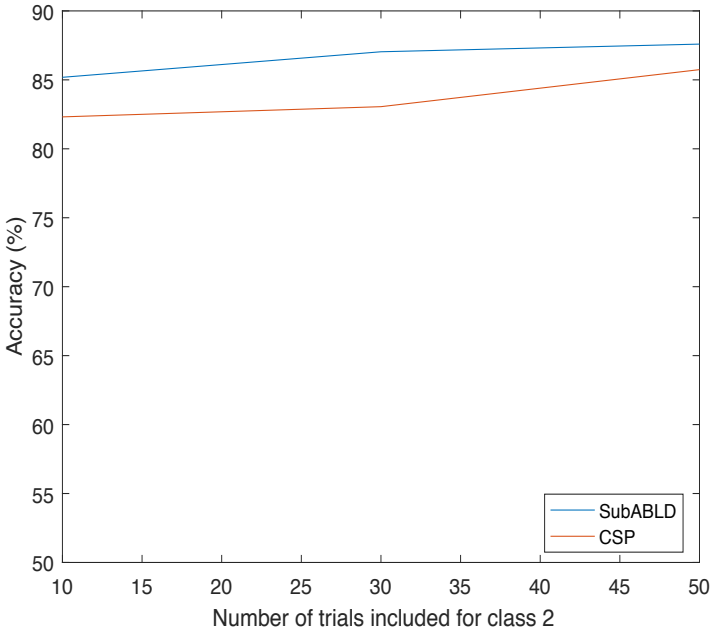


(a)

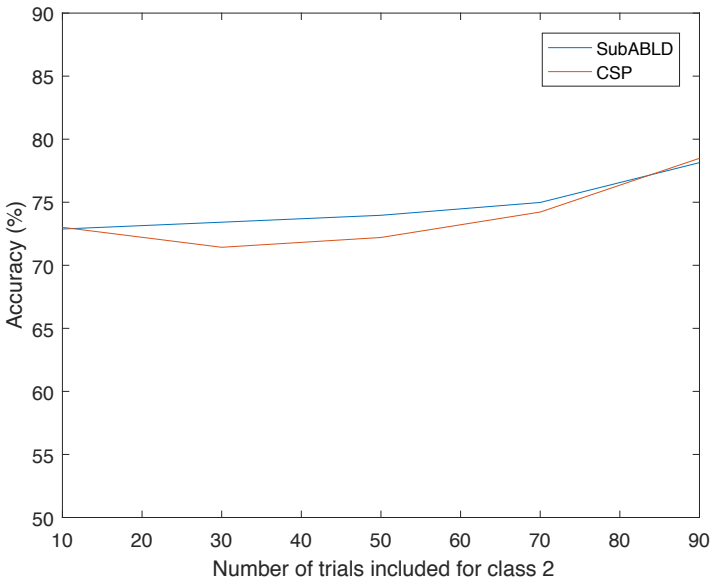


(b)

Fig. 8.12 (a) Performance comparison of Sub-ABLD algorithm ($\eta = 2,\ \alpha = \beta = 1.5$) and CSP by increasing the number of outlier trials using BCI competition datasets III dataset 3a and (b)Performance comparison of Sub-ABLD algorithm ($\eta = 0.25,\ \alpha = \beta = 1.25$) and CSP by increasing the number of outlier trials using BCI competition datasets IV dataset 2a.

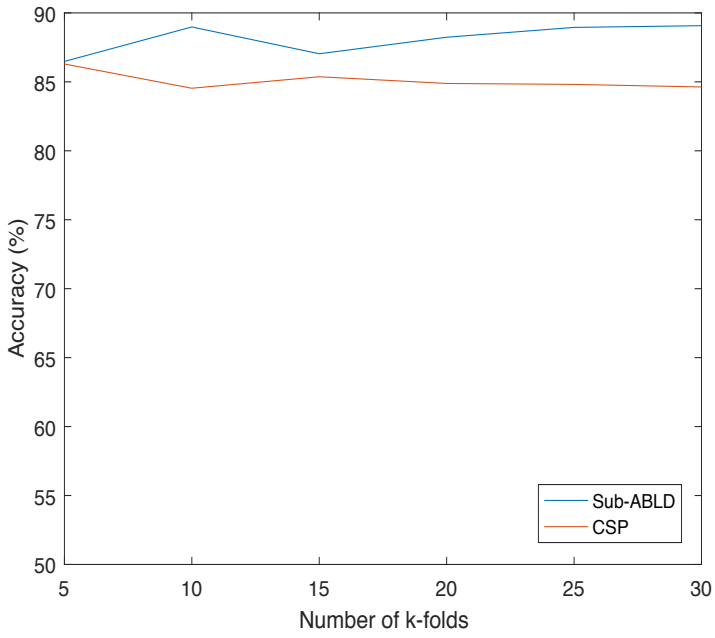


(a)

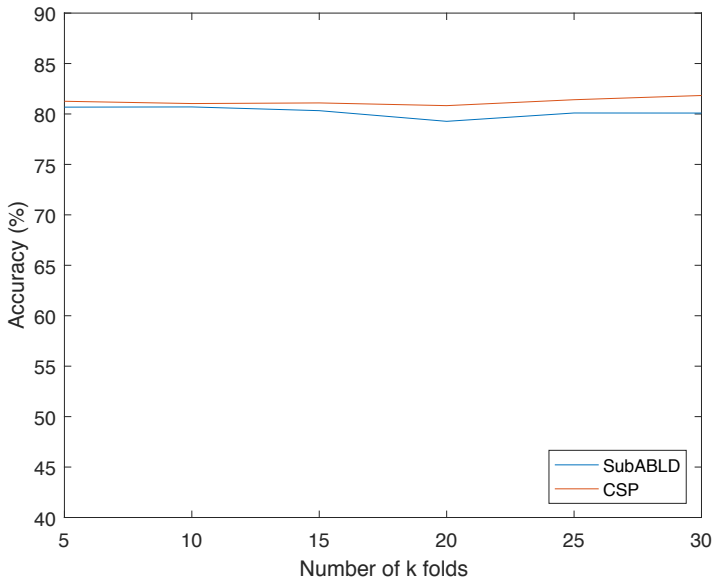


(b)

Fig. 8.13 (a) Performance comparison of Sub-ABLD algorithm ($\eta = 2,\ \alpha = \beta = 1.5$) and CSP by imbalancing the number of training trials for two class using BCI competition datasets III dataset 3a and (b) Performance comparison Sub-ABLD algorithm ($\eta = 0.25,\ \alpha = \beta = 1.25$) and CSP by imbalancing the number of training trials for two class using BCI competition datasets IV dataset 2a.



(a)



(b)

Fig. 8.14 (a) Performance comparison of Sub-ABLD algorithm ($\eta = 2,\ \alpha = \beta = 1.5$) and CSP by varying the number of training trials using BCI competition datasets III dataset 3a and (b) Performance comparison Sub-ABLD algorithm ($\eta = 0.25,\ \alpha = \beta = 1.25$) and CSP by varying the number of training trials using BCI competition datasets IV dataset 2a.

The next experiment is done by imbalancing the number of trials used in the training set. The results obtained during this experiment are shown in Fig. 8.13. It shows that the performance is poor if the number of trials used for computing the spatial filters from

the two classes is unequal. Hence, we observed the better performance by increasing the number of trials equally towards class 1 for both the algorithms. This indicates that the algorithms perform better by selecting the same number of trials for both the class. Finally, the performance results presented in Fig. 8.14 are obtained by varying the number of training trials. The observed figures show that the overall performance when k = 10 and k=30 is the same, but the time taken for simulation is more for k=30. Hence, 10 folds cv is appropriate for these two datasets.

8.3 Conclusions

In this chapter, different studies have been performed to analyze the robustness of the proposed algorithms. The performance and pattern of ThinICA-CSP is compared with mCSP and JADE. It shows that ThinICA-CSP outperforms the other two algorithms. The Sub-ABLD algorithm was tested with both artificial and real data and the performance comparison with the other existing algorithms like CSP, JADE, MAP-CSP and divCSP-WS was also presented. Sub-ABLD outperforms the other algorithm in the artificial and BCI competition III dataset 3a whereas the performance is same with CSP and JADE for the other two datasets. Furthermore, we have also studied the performance of the algorithm in various scenarios.

CHAPTER 9

Conclusions

This thesis presents two algorithms for the classification of motor imagery movements in BCI applications. At the start, the anatomy and physiology of the human brain, EEG signals, different brain rhythms and artifacts present in the EEG signals were introduced. The MI-based BCI system and steps involved in it such as filtering and classification techniques were also discussed. Later, various existing spatial filtering algorithms and the challenges are studied. Based on this study, two spatial filtering algorithms were proposed.

The first algorithm is based on the extension of ThinICA and mCSP to address the problem for discrimination of four class motor imagery movements. In this approach, the contrast function is obtained by combining the second order and higher-order statistics. The existing ThinICA algorithm extracts the independent components by considering only the higher order statistic, but for the non-stationary sources, it is more convenient to analyze the data by splitting it into K blocks. Thus, the marginal contrast function was evaluated for each split and simultaneously maximize the accumulated sum of the marginal contrasts. Furthermore, the combination of the several lower-order statistics of the outputs with delays was also incorporated in the contrast function to estimate the demixing matrix. Although ICA can recover the subset of independent components, it is critical to select only the MI related components. Therefore, it can be done by initializing the unmixing matrix with the solution provided by mCSP. The comparative study using LDA and SVM classifiers was also performed. It is observed that LDA performs better than SVM. The overall performance result shows that the proposed algorithm performs better than mCSP and JADE. This indicates that the utilization of second and higher order statistics and initialization of mixing matrix with the mCSP solutions improves the classification performance. Hence, this provides an additional evidence in favor of ICA techniques for designing a robust classification algorithm.

The second algorithm is based on AB Log-Det divergence, which is the main contribution of this thesis. The AB Log-Det divergence optimization problem has been interpreted in terms of CSP criterion. The κ parameter that provides equivalent solutions between the CSP and AB Log-Det divergence optimization was obtained. The gradient for AB Log-Det divergence was derived. The robustness of the criterion with respect to

the hyperparameters α and β was also analyzed. Based on this criterion, the Sub-ABLD algorithm is proposed to address the problem of discrimination between two classes. The optimization is performed by considering both the within class and between class divergences. The testing of the algorithm is done using artificial as well as real datasets. Different studies were performed to test the robustness of the algorithm by including the outliers trial, varying the number of training trials, imbalancing the training trials for two class and using different averaging techniques. Moreover, the comparison of the performance of the proposed algorithm with the other existing algorithms was also done. The observed results indicate that the proposed algorithm outperforms the other algorithms for the simulated and BCI competition III dataset 3a. For BCI competition III dataset IVa and Competition IV dataset 2a, the proposed algorithm performs equally with the other algorithms like JADE, divCSP-WS and CSP. The study with Arithmetic and Geometric mean shows that the Arithmetic mean gives better results than the Geometric mean for BCI competition III dataset 3a. However, both the methods performed equally for BCI competition IV dataset 2a. From this study, it can also be concluded that selecting an equal number of training trials for both the class gives better performance. The overall results show the robustness of the proposed Sub-ABLD algorithm.

9.1 Future Work

Two robust algorithms for the discrimination of MI movements in BCI applications have been proposed. The obtained results outperform the baseline methods but still needs some tuning and further studies.

The ICA techniques extract independent components but it is still a challenging task to extract the motor imaginary related components. Hence, obtaining reference signals that are closely related to motor imagery movements (or artifacts) and selecting the components based on the maximum (or minimum) mutual information between the reference signals and independent components may improve the performance. Moreover, a regularized term which is computed using artifacts or related signal can be incorporated into the ThinICA-CSP objective function to improve the performance.

The second algorithm of this thesis is based on AB Log-Det divergence. It provides a generalized form that can derive other types of dissimilarity measures like squared Riemannian metric, the Steins loss, the S-divergence by varying α and β parameters. Further, the regularization parameter η also influenced the algorithm performance. Therefore, selecting appropriate hyperparameters always play a major role in the optimization problem. The common way for selecting these parameters is using cross-validation process. Hence, a proper selection method and proper range of parameters can be defined for improving the performance.

Moreover, after achieving the acceptable classification accuracy for practical use, it

can be used for the real time application such as interfacing with assistive devices like wheelchair etc. to assist the paralysed subject.

REFERENCES

Allwein, E. L., Schapire, R. E. and Singer, Y. (2001), 'Reducing multiclass to binary: A unifying approach for margin classifiers', *The Journal of Machine Learning Research* **1**, 113–141.

Amari, S. I. (1998), 'Natural gradient works efficiently in learning', *Neural Computation* **10**(2), 251–276.

Amari, S. I., Cichocki, A. and Yang, H. H. (1996), A new learning algorithm for blind signal separation, *in* 'Advances in Neural Information Processing Systems', pp. 757–763.

Ang, K. K., Chin, Z. Y., Zhang, H. and Guan, C. (2008), Filter bank common spatial pattern (FBCSP) in brain-computer interface, *in* 'IEEE International Joint Conference on Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence)', IEEE, pp. 2390–2397.

Arvaneh, M., Guan, C., Ang, K. K. and Quek, C. (2010), Optimizing EEG channel selection by regularized spatial filtering and multi band signal decomposition, *in* 'IASTED International Conference on Biomedical Engineering', pp. 86–90.

Arvaneh, M., Guan, C., Ang, K. K. and Quek, C. (2011*a*), 'Optimizing the channel selection and classification accuracy in EEG-based BCI', *IEEE Transactions on Biomedical Engineering* **58**(6), 1865–1873.

Arvaneh, M., Guan, C., Ang, K. K. and Quek, C. (2013), 'Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface', *IEEE Transactions on Neural Networks and Learning Systems* **24**(4), 610–619.

Arvaneh, M., Guan, C., Ang, K. K. and Quek, H. C. (2011*b*), Spatially sparsed common spatial pattern to improve BCI performance, *in* '2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 2412–2415.

Barachant, A., Bonnet, S., Congedo, M. and Jutten, C. (2012), 'Multiclass brain–computer interface classification by Riemannian geometry', *IEEE Transactions on Biomedical Engineering* **59**(4), 920–928.

Barber, D. (2012), *Bayesian Reasoning and Machine Learning*, Cambridge University Press.

BCI Competition III (2005).

BCI Competition IV (2008).

Bell, A. J. and Sejnowski, T. J. (1995), 'An information-maximization approach to blind separation and blind deconvolution', *Neural Computation* **7**(6), 1129–1159.

Belouchrani, A., Abed Meraim, K., Cardoso, J. F. and Moulines, E. (1997), 'A blind source separation technique using second-order statistics', *IEEE Transactions on Signal Processing* **45**(2), 434–444.

Berger, H. (1929), 'Über das elektrenkephalogramm des menschen', *European Archives of Psychiatry and Clinical Neuroscience* **87**(1), 527–570.

Bhatia, R. (1997), 'Matrix analysis graduate texts in mathematics, 169'.

Bingham, E. and Hyvärinen, A. (2000), 'A fast fixed-point algorithm for independent component analysis of complex valued signals', *International Journal of Neural Systems* **10**(01), 1–8.

Blankertz, B., Curio, G. and Müller, K.-R. (2002), Classifying single trial eeg: Towards brain computer interfacing, *in* 'Advances in neural information processing systems', pp. 157–164.

Blankertz, B., Kawanabe, M., Tomioka, R., Hohlefeld, F., Müller, K. R. and Nikulin, V. V. (2007), Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing, *in* 'Advances in Neural Information Processing Systems', pp. 113–120.

Blankertz, B., Müller, K. R., Krusienski, D. J., Schalk, G., Wolpaw, J. R., Schlögl, A., Pfurtscheller, G., Millan, J. R., Schröder, M. and Birbaumer, N. (2006), 'The BCI competition III:Validating alternative approaches to actual BCI problems', *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **14**(2), 153–159.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M. and Muller, K. R. (2008), 'Optimizing spatial filters for robust EEG single-trial analysis', *IEEE Signal Processing Magazine* **25**(1), 41–56.

Blume, W. T. (1999), 'Atlas of pediatric electroencephalography'.

Bonnet, L., Lotte, F. and Lécuyer, A. (2013), 'Two brains, one game: design and evaluation of a multiuser BCI video game based on motor imagery', *IEEE Transactions on Computational Intelligence and AI in games* **5**(2), 185–198.

Brandl, S., Müller, K. R. and Samek, W. (2015), Robust common spatial patterns based on Bhattacharyya distance and Gamma divergence, *in* '2015 3rd International Winter Conference on Brain-Computer Interface (BCI)', IEEE, pp. 1–4.

Brunner, C., Leeb, R., Müller-Putz, G., Schlögl, A. and Pfurtscheller, G. (2008), 'BCI Competition 2008–Graz data set A', *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology* p. 16.

Brunner, C., Naeem, M., Leeb, R., Graimann, B. and Pfurtscheller, G. (2007), 'Spatial filtering and selection of optimized components in four class motor imagery EEG data using independent components analysis', *Pattern Recognition Letters* **28**(8), 957–964.

Chin, Z. Y., Ang, K. K., Wang, C., Guan, C. and Zhang, H. (2009), Multi-class filter bank common spatial pattern for four-class motor imagery BCI, *in* 'Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009.', IEEE, pp. 571–574.

Chiou, J.-C., Ko, L.-W., Lin, C.-T., Hong, C.-T., Jung, T.-P., Liang, S.-F. and Jeng, J.-L. (2006), Using novel mems eeg sensors in detecting drowsiness application, *in* 'Biomedical Circuits and Systems Conference, 2006. BioCAS 2006. IEEE', IEEE, pp. 33–36.

Cichocki, A.; Amari, S. (2010), 'Families of Alpha- Beta- and Gamma- divergences: Flexible and robust measures of similarities', *Entropy* **12**, 1532–1568.

Cichocki, A., Cruces, S. and Amari, S. I. (2015), 'Log-determinant divergences revisited: Alpha-Beta and Gamma log-det divergences', *Entropy* **17**(5), 2988–3034.

Cortes, C. and Vapnik, V. (1995), 'Support-vector networks', *Machine Learning* **20**(3), 273–297.

Cruces-Alvarez, S. A., Cichocki, A. and Amari, S. I. (2004), 'From blind signal extraction to blind instantaneous signal separation: criteria, algorithms, and stability', *IEEE Transactions on Neural Networks* **15**(4), 859–873.

Cruces, S. and Cichocki, A. (2003), Combining blind source extraction with joint approximate diagonalization: Thin algorithms for ICA, *in* 'International Symposium on Independent Component Analysis and Blind Signal Separation, Japan', IEEE, pp. 463–468.

Cruces, S., Cichocki, A. and Amari, S. I. (2004), 'From blind signal extraction to blind instantaneous signal separation', *IEEE Transactions on Neural Networks* **15**(4), 859–873.

Cruces, S., Cichocki, A. and De Lathauwer, L. (2004), Thin QR and SVD factorizations for simultaneous blind signal extraction, *in* '2004 12th European Signal Processing Conference,', IEEE, pp. 217–220.

Dornhege, G. (2007), *Toward Brain-Computer Interfacing*, MIT press.

Dornhege, G., Blankertz, B., Curio, G. and Müller, K. R. (2004), 'Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms', *IEEE Transactions on Biomedical Engineering* **51**(6), 993–1002.

Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G. and Muller, K. R. (2006), 'Combined optimization of spatial and temporal filters for improving brain-computer interfacing', *IEEE Transactions on Biomedical Engineering* **53**(11), 2274–2281.

Duda, R. O. and Hart, P. E. (1973), *Pattern Classification and Scene Analysis*, Vol. 3, Wiley New York.

Edelman, A., Arias, T. A. and Smith, S. T. (1998), 'The geometry of algorithms with orthogonality constraints', *SIAM Journal on Matrix Analysis and Applications* **20**(2), 303–353.

Farquhar, J., Hill, N., Lal, T. N. and Schölkopf, B. (2006), Regularised CSP for sensor selection in BCI, *in* '3rd International BCI workshop, Austria', pp. 1–2.

Feder, M. and Merhav, N. (1994), 'Relations between entropy and error probability', *IEEE Transactions on Information Theory* **40**(1), 259–266.

Fukunaga, K. and KoonTz, W. L. G. (1970), 'Application of the Karhunen-Loeve expansion to feature selection and ordering', *IEEE Transactions on Computers* **C-19**(4), 440–447.

Gargiulo, G., Calvo, R. A., Bifulco, P., Cesarelli, M., Jin, C., Mohamed, A. and van Schaik, A. (2010), 'A new eeg recording system for passive dry electrodes', *Clinical Neurophysiology* **121**(5), 686–693.

Ge, S., Wang, R. and Yu, D. (2014), 'Classification of four-class motor imagery employing single-channel electroencephalography', *PloS one* **9**(6), e98019.

Gouy-Pailler, C., Congedo, M., Brunner, C., Jutten, C. and Pfurtscheller, G. (2010), 'Nonstationary brain source separation for multiclass motor imagery', *IEEE Transactions on Biomedical Engineering* **57**(2), 469–478.

Griss, P., Tolvanen-Laakso, H. K., Merilainen, P. and Stemme, G. (2002), 'Characterization of micromachined spiked biopotential electrodes', *IEEE Transactions on Biomedical Engineering* **49**(6), 597–604.

Grosse-Wentrup, M. and Buss, M. (2008), 'Multiclass common spatial patterns and information theoretic feature extraction', *IEEE Transactions on Biomedical Engineering* **55**(8), 1991–2000.

Grosse-Wentrup, M., Liefhold, C., Gramann, K. and Buss, M. (2009), 'Beamforming in noninvasive brain–computer interfaces', *IEEE Transactions on Biomedical Engineering* **56**(4), 1209–1219.

Grozea, C., Voinescu, C. D. and Fazli, S. (2011), 'Bristle-sensorslow-cost flexible passive dry eeg electrodes for neurofeedback and bci applications', *Journal of neural engineering* **8**(2), 025008.

Guermeur, Y. and Monfrini, E. (2011), 'A quadratic loss multi-class SVM for which a radius–margin bound applies', *Informatica* **22**(1), 73–96.

Harandi, M., Salzmann, M. and Hartley, R. (2017), 'Dimensionality reduction on SPD manifolds: The emergence of geometry-aware methods', *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Harland, C., Clark, T. and Prance, R. (2002), 'Remote detection of human electroencephalograms using ultrahigh input impedance electric potential sensors', *Applied Physics Letters* **81**(17), 3284–3286.

Hiraiwa, A., Shimohara, K. and Tokunaga, Y. (1990), 'EEG topography recognition by neural networks', *IEEE Engineering in Medicine and Biology magazine* **9**(3), 39–42.

Horev, I., Yger, F. and Sugiyama, M. (2016), Geometry-aware principal component analysis for symmetric positive definite matrices, *in* 'Asian Conference on Machine Learning', pp. 1–16.

Hotelling, H. (1933), 'Analysis of a complex of statistical variables into principal components.', *Journal of Educational Psychology* **24**(6), 417.

Huang, Z., Wang, R., Shan, S., Li, X. and Chen, X. (2015), Log-Euclidean metric learning on symmetric positive definite manifold with application to image set classification, *in* 'International Conference on Machine Learning', pp. 720–729.

Hyvärinen, A., Karhunen, J. and Oja, E. (2004), *Independent component analysis*, Vol. 46, John Wiley and Sons.

Hyvärinen, A. and Oja, E. (1996), A neuron that learns to separate one independent component from linear mixtures, *in* 'Proceedings IEEE International Conference on Neural Networks. Washington, DC', pp. 62–67.

Jasper, H. (1958), 'Report of the committee on methods of clinical examination in electroencephalography', *Electroencephalography and Clinical Neurophysiology* **10**, 370–375.

Jones, M. C. and Sibson, R. (1987), 'What is projection pursuit?', *Journal of the Royal Statistical Society. Series A (General)* pp. 1–37.

Jung, T. P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E. and Sejnowski, T. J. (2000), 'Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects', *Clinical Neurophysiology* **111**(10), 1745–1758.

Kang, H., Nam, Y. and Choi, S. (2009), 'Composite common spatial pattern for subject-to-subject transfer', *Signal Processing Letters, IEEE* **16**(8), 683–686.

Kawanabe, M., Samek, W., Müller, K. R. and Vidaurre, C. (2014), 'Robust common spatial filters with a maxmin approach', *Neural Computation* **26**(2), 349–376.

Kawanabe, M. and Vidaurre, C. (2009), Improving BCI performance by modified common spatial patterns with robustly averaged covariance matrices, *in* 'World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany', Springer, pp. 279–282.

Kawanabe, M., Vidaurre, C., Scholler, S. and Müller, K. R. (2009), Robust common spatial filters with a maxmin approach, *in* '2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society', IEEE, pp. 2470–2473.

Koles, Z. J. (1991), 'The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG', *Electroencephalography and Clinical Neurophysiology* **79**(6), 440–447.

Koles, Z., Lind, J. and Flor-Henry, P. (1994), 'Spatial patterns in the background EEG underlying mental disease in man', *Electroencephalography and Clinical Neurophysiology* **91**(5), 319–328.

Krauledat, M., Grzeska, K., Sagebaum, M., Blankertz, B., Vidaurre, C., Müller, K.-R. and Schröder, M. (2009), Playing pinball with non-invasive BCI, *in* 'Advances in Neural Information Processing Systems', pp. 1641–1648.

Lauer, F. and Guermeur, Y. (2011), 'MSVMpack: a multi-class support vector machine package', *Journal of Machine Learning Research* **12**(Jul), 2293–2296.

Lemm, S., Blankertz, B., Curio, G. and Muller, K. R. (2005), 'Spatio-spectral filters for improving the classification of single trial EEG', *IEEE Transactions on Biomedical Engineering* **52**(9), 1541–1548.

Li, R. (2013), 'Rayleigh quotient based optimization methods for eigenvalue problems', *In Summary of Lectures Delivered at Gene Golub SIAM Summer School 2013, Fudan University: Shanghai, China* pp. 1–27.

Liao, L.-D., Wang, I.-J., Chen, S.-F., Chang, J.-Y. and Lin, C.-T. (2011), 'Design, fabrication and experimental validation of a novel dry-contact sensor for measuring electroencephalography signals without skin preparation', *Sensors* **11**(6), 5819–5834.

Llera, A., Gómez, V. and Kappen, H. J. (2014), 'Adaptive multiclass classification for brain computer interfaces', *Neural Computation* **26**(6), 1108–1127.

Lotte, F. and Guan, C. (2010*a*), Learning from other subjects helps reducing brain-computer interface calibration time, *in* '2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)', IEEE, pp. 614–617.

Lotte, F. and Guan, C. (2010*b*), Spatially regularized common spatial patterns for EEG classification, *in* '2010 20th International Conference on Pattern Recognition (ICPR)', IEEE, pp. 3712–3715.

Lotte, F. and Guan, C. (2011), 'Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms', *IEEE Transactions on Biomedical Engineering* **58**(2), 355–362.

Lu, H., Eng, H. L., Guan, C., Plataniotis, K. N. and Venetsanopoulos, A. N. (2010), 'Regularized common spatial pattern with aggregation for EEG classification in small-sample setting', *IEEE Transactions on Biomedical Engineering* **57**(12), 2936–2946.

Lu, H., Plataniotis, K. N. and Venetsanopoulos, A. N. (2009), Regularized common spatial patterns with generic learning for EEG signal classification, *in* '2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society', IEEE, pp. 6599–6602.

Machine Learning in Neural Engineering (2017(last update))).

Makeig, S., Debener, S., Onton, J. and Delorme, A. (2004), 'Mining event-related brain dynamics', *Trends in Cognitive Sciences* **8**(5), 204–210.

Müller-Gerking, J., Pfurtscheller, G. and Flyvbjerg, H. (1999), 'Designing optimal spatial filters for single-trial EEG classification in a movement task', *Clinical Neurophysiology* **110**(5), 787–798.

Naeem, M., Brunner, C., Leeb, R., Graimann, B. and Pfurtscheller, G. (2006), 'Seperability of four-class motor imagery data using independent components analysis', *Journal of Neural Engineering* **3**(3), 208.

Nguyen, T. H., Park, S. M., Ko, K. E. and Sim, K. B. (2012), Multi-class stationary CSP for optimal feature separation of brain source in BCI system, *in* '2012 12th International Conference on Control, Automation and Systems (ICCAS)', IEEE, pp. 1035–1039.

Nicolas-Alonso, L. F., Corralejo, R., Gomez-Pilar, J., Álvarez, D. and Hornero, R. (2015), 'Adaptive stacked generalization for multiclass motor imagery-based brain computer interfaces', *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **23**(4), 702–712.

Nishimori, Y. (1999), Learning algorithm for independent component analysis by geodesic flows on orthogonal group, *in* 'International Joint Conference on Neural Networks, 1999. IJCNN'99', Vol. 2, IEEE, pp. 933–938.

Oehler, M., Neumann, P., Becker, M., Curio, G. and Schilling, M. (2008), Extraction of ssvep signals of a capacitive eeg helmet for human machine interface, *in* 'Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE', IEEE, pp. 4495–4498.

Park, J. and Chung, W. (2013), Common spatial patterns based on generalized norms, *in* '2013 International Winter Workshop on Brain-Computer Interface (BCI)', IEEE, pp. 39–42.

Pfurtscheller, G. (1999), 'EEG event-related desynchronization (ERD) and event-related sybchronization (ERS)', *Electroencephalography: Basic principals, clinical applications and related fields* pp. 958–967.

Pfurtscheller, G., Brunner, C., Schlögl, A. and Da Silva, F. L. (2006), 'Mu rhythm (de) synchronization and EEG single-trial classification of different motor imagery tasks', *NeuroImage* **31**(1), 153–159.

Pfurtscheller, G. and Da Silva, F. L. (1999), 'Event-related EEG/MEG synchronization and desynchronization: basic principles', *Clinical Neurophysiology* **110**(11), 1842–1857.

Pfurtscheller, G., Flotzinger, D. and Kalcher, J. (1993), 'Brain-computer interfacea new communication device for handicapped persons', *Journal of Microcomputer Applications* **16**(3), 293–299.

Quang, M. H. (2016), 'Infinite-dimensional Log-Determinant divergences II: Alpha-Beta divergences', *arXiv preprint arXiv:1610.08087* .

Ramoser, H., Müller-Gerking, J. and Pfurtscheller, G. (2000), 'Optimal spatial filtering of single trial EEG during imagined hand movement', *IEEE Transactions on Rehabilitation Engineering* **8**(4), 441–446.

Rezaei, S., Tavakolian, K., Nasrabadi, A. M. and Setarehdan, S. K. (2006), 'Different classification techniques considering brain computer interface applications', *Journal of Neural Engineering* **3**(2), 139.

Ruffini, G., Dunne, S., Fuentemilla, L., Grau, C., Farres, E., Marco-Pallarés, J., Watts, P. and Silva, S. (2008), 'First human trials of a dry electrophysiology sensor using a carbon nanotube array interface', *Sensors and Actuators A: Physical* **144**(2), 275–279.

Salakhutdinov, R. and Mnih, A. (2008), Bayesian probabilistic matrix factorization using markov chain monte carlo, *in* 'Proceedings of the 25th international conference on Machine learning', ACM, pp. 880–887.

Salvo, P., Raedt, R., Carrette, E., Schaubroeck, D., Vanfleteren, J. and Cardon, L. (2012), 'A 3d printed dry electrode for ecg/eeg recording', *Sensors and Actuators A: Physical* **174**, 96–102.

Samek, W., Binder, A. and Müller, K. R. (2013), Multiple kernel learning for brain-computer interfacing, *in* '2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)', IEEE, pp. 7048–7051.

Samek, W., Blythe, D., Müller, K. R. and Kawanabe, M. (2013), Robust spatial filtering with beta divergence, *in* 'Advances in Neural Information Processing Systems', pp. 1007–1015.

Samek, W., Kawanabe, M. and Müller, K. R. (2014), 'Divergence-based framework for common spatial patterns algorithms', *IEEE Reviews in Biomedical Engineering* **7**, 50–72.

Samek, W., Kawanabe, M. and Vidaurre, C. (2011), Group-wise stationary subspace analysis–a novel method for studying non-stationarities, *in* 'Proceedings International Brain–Computer Interfacing Conference', pp. 16–20.

Samek, W. and Müller, K. R. (2015), Tackling noise, artifacts and nonstationarity in BCI with robust divergences, *in* '2015 23rd European Signal Processing Conference (EUSIPCO)', IEEE, pp. 2741–2745.

Samek, W., Müller, K.-R., Kawanabe, M. and Vidaurre, C. (2012), Brain-computer interfacing in discriminative and stationary subspaces, *in* '2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)', IEEE, pp. 2873–2876.

Samek, W., Vidaurre, C., Müller, K. R. and Kawanabe, M. (2012), 'Stationary common spatial patterns for brain–computer interfacing', *Journal of Neural Engineering* **9**(2), 026013.

Schlögl, A., Lee, F., Bischof, H. and Pfurtscheller, G. (2005), 'Characterization of four-class motor imagery EEG data for the BCI-competition 2005', *Journal of Neural Engineering* **2**(4), L14.

Spilker, B., Kamiya, J., Callaway, E. and Yeager, C. L. (1969), 'Visual evoked responses in subjects trained to control alpha rhythms', *Psychophysiology* **5**(6), 683–695.

Sullivan, T. J., Deiss, S. R. and Cauwenberghs, G. (2007), A low-noise, non-contact eeg/ecg sensor, *in* 'Biomedical Circuits and Systems Conference, 2007. BIOCAS 2007. IEEE', IEEE, pp. 154–157.

Tangermann, M., Müller, K. E., clinical neurophysiologyRobert, Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K. J., Mueller-Putz, G. et al. (2012), 'Review of the BCI competition IV', *Frontiers in Neuroscience* **6**, 55.

Tao, T. (2012), *Topics in Random Matrix Theory*, Vol. 132, American Mathematical Soc.

The Divergence Methods Web Site (2013).

Townsend, G., Graimann, B. and Pfurtscheller, G. (2006), 'A comparison of common spatial patterns with complex band power features in a four-class BCI experiment', *IEEE Transactions on Biomedical Engineering* **53**(4), 642–651.

Urigüen, J. A. and Garcia-Zapirain, B. (2015), 'EEG artifact removal-state-of-the-art and guidelines', *Journal of Neural Engineering* **12**(3), 031001.

Van Erp, J., Lotte, F. and Tangermann, M. (2012), 'Brain-computer interfaces: beyond medical applications', *Computer* **45**(4), 26–34.

Vidal, J. J. (1973), 'Toward direct brain-computer communication', *Annual review of Biophysics and Bioengineering* **2**(1), 157–180.

Von Bünau, P., Meinecke, F. C., Király, F. C. and Müller, K. R. (2009), 'Finding stationary subspaces in multivariate time series', *Physical Review Letters* **103**(21), 214101.

Von Bünau, P., Meinecke, F. C., Scholler, S. and Müller, K. R. (2010), Finding stationary brain sources in EEG data, *in* '2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)', IEEE, pp. 2810–2813.

Wang, H. (2012), 'Harmonic mean of Kullback–Leibler divergences for optimizing multi-class EEG spatio-temporal filters', *Neural Processing Letters* **36**(2), 161–171.

Wang, H. and Li, X. (2016), 'Regularized filters for L1-norm-based common spatial patterns', *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **24**(2), 201–211.

Wang, H., Tang, Q. and Zheng, W. (2012), 'L1-norm-based common spatial patterns', *IEEE Transactions on Biomedical Engineering* **59**(3), 653–662.

Wojcikiewicz, W., Vidaurre, C. and Kawanabe, M. (2011*a*), Improving classification performance of BCIs by using stationary common spatial patterns and unsupervised bias adaptation, *in* 'International Conference on Hybrid Artificial Intelligence Systems', Springer, pp. 34–41.

Wojcikiewicz, W., Vidaurre, C. and Kawanabe, M. (2011*b*), Stationary common spatial patterns: towards robust classification of non-stationary EEG signals, *in* '2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 577–580.

Wolpaw, J. and Wolpaw, E. W. (2012), *Brain-computer Interfaces: Principles and Practice*, Oxford University Press.

Wu, W., Chen, Z., Gao, X., Li, Y., Brown, E. N. and Gao, S. (2015), 'Probabilistic common spatial patterns for multichannel EEG analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**(3), 639–653.

Xinyi Yong, R. K. W. and Birch, G. E. (2008), Robust common spatial patterns for EEG signal preprocessing, *in* 'IEEE EMBS 30th Annual International Conference', IEEE, pp. 2087– 2090.

Xu, P., Yang, P., Lei, X. and Yao, D. (2011), 'An enhanced probabilistic LDA for multi-class brain computer interface', *PloS one* **6**(1), e14634.

Yong, X., Ward, R. K. and Birch, G. E. (2008*a*), Robust common spatial patterns for EEG signal preprocessing, *in* '30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008. EMBS 2008.', IEEE, pp. 2087–2090.

Yong, X., Ward, R. K. and Birch, G. E. (2008*b*), Sparse spatial filter optimization for EEG channel reduction in brain-computer interface, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008', IEEE, pp. 417–420.

Zhang, H., Yang, H. and Guan, C. (2013), 'Bayesian learning for spatial filtering in an EEG-based brain–computer interface', *IEEE Transactions on Neural Networks and Learning Systems* **24**(7), 1049–1060.

Zhou, B., Wu, X., Zhang, L., Lv, Z. and Guo, X. (2014), 'Robust spatial filters on three-class motor imagery EEG data using independent component analysis', *Journal of Biosciences and Medicines* **2**(02), 43.

LIST OF PUBLICATIONS

1. Deepa Beeta, Th., Sergio, C., Javier, O., Andrzej C.,2017. Optimization of Aplha-Beta Log-Det Divergences and their Applications in the Spatial Filtering of Two Class Motor Imagery Movements. Entropy ,19(3): 89, pp. 1-40.

2. Deepa Beeta, Th., Rajkumar E.R., 2016. A Comparative Performance Analysis for Classification of Multiclass Motor Imagery Movements. International Journal of Control Theory and Applications, 9(36), pp. 443-450.

3. Deepa Beeta, Th., Rajkumar E.R.,2016. Common Spatial Pattern Algorithm Based Signal Processing Techniques for Classification of Motor Imagery Movements: A Mini Review. International Journal of Control Theory and Applications (Accepted).

4. Deepa Beeta, Th., Sergio, C., Rajkumar E.R., 2016. ThinICA-CSP algorithm for discrimination of multiclass motor imagery movements. In Region 10 Conference (TENCON), IEEE,pp. 2483-2486.

5. Pablo, A.,Deepa Beeta, Th., Auxiliadora, S.,Irene, F., 2014. Cancelacin de interferencias en EEG mediante el Anlisis de Componentes Acotadas. In Congreso Anual de la Sociedad Espaola de Ingeniera Biomdica (CASEIB), pp. 1-4.

6. Deepa Beeta, Th., Rajkumar E.R., 2014. Development of Cost Effective Backscattered Optical Imaging System for Detection of Abnormality in Tissues. In IEEE Engineering in Medicine and Biology Society (EMB), IEEE. (Poster).

Appendices

Appendix A

DETERMINATION OF THE UPPER-BOUND OF THE AB-LOG-DET-DIVERGENCE

The divergence $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ depends on the generalized eigenvalues of the matrix pencil $(\mathbf{P},\mathbf{Q})$, which have been denoted by λ_i for $i = 1,\ldots,n$. Similarly, the divergence of the compressed arguments $D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\|\ \mathbf{W}^T\mathbf{Q}\mathbf{W})$ depends on μ_i for $i = 1,\ldots,p$, the eigenvalues of the matrix pencil $(\mathbf{W}^T\mathbf{P}\mathbf{W},\mathbf{W}^T\mathbf{Q}\mathbf{W})$. The Cauchy interlacing inequalities Li (2013)

$$\lambda_j \leq \mu_j \leq \lambda_{n-p+j} \tag{A1}$$

provide upper and lower-bounds for μ_j in terms of the eigenvalues of the uncompressed matrix pencil. This property implies that the eigenvalues μ_j, for each $j = 1,\ldots,p$, should lie in a sequence of possibly partially overlapping intervals given by $[\lambda_j,\lambda_{n-p+j}]$.

The divergence $D_{AB}^{(\alpha,\beta)}(\lambda\|\ 1)$ is minimum (zero) for $\lambda = 1$, strictly monotone descending for $\lambda < 1$ and strictly monotone ascending for $\lambda > 1$. So we can bound the the AB log-det divergence in each interval by

$$D_{AB}^{(\alpha,\beta)}(\mu_j\|\ 1) \ \leq\ \max\{D_{AB}^{(\alpha,\beta)}(\lambda_j\|\ 1)\ ,\ D_{AB}^{(\alpha,\beta)}(\lambda_{n-p+j}\|\ 1)\}, \tag{A2}$$

and the maximum value occurs at one of the extreme eigenvalues of the interval. The construction of the interlacing property, prevents that any eigenvalue with a given index could appear more than once as upper extreme of an interval or as a lower extreme of an interval. This fact, combined with the strict monotonicity property of the divergence, implies that the maxima of the divergence for each interval can only be obtained by eigenvalues with different indices. Finally, the result of adding these p maximum values can not exceed the sum of the divergences for those eigenvalues which maximize the divergence from unity,

$$D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W}\ \|\mathbf{W}^T\mathbf{Q}\mathbf{W}) = \sum_{j=1}^{p} D_{AB}^{(\alpha,\beta)}(\mu_j\|\ 1) \tag{A3}$$

$$\leq \sum_{j=1}^{p} \max\{D_{AB}^{(\alpha,\beta)}(\lambda_j\|\ 1), D_{AB}^{(\alpha,\beta)}(\lambda_{n-p+j}\|\ 1)\} \tag{A4}$$

With the help of the permutation π of the indices $1, \ldots, n$ that sorts the divergence of the eigenvalues from the unity in descending order

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_1} \| \, 1) \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_2} \| \, 1) \geq \cdots \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_n} \| \, 1)\,, \tag{A5}$$

we can write

$$D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{P}\mathbf{W} \, \| \mathbf{W}^T\mathbf{Q}\mathbf{W}) = \sum_{j=1}^{p} D_{AB}^{(\alpha,\beta)}(\mu_j \| \, 1) \tag{A6}$$

$$\leq \sum_{j=1}^{p} \max\{D_{AB}^{(\alpha,\beta)}(\lambda_j \| \, 1) \; , \; D_{AB}^{(\alpha,\beta)}(\lambda_{n-p+j} \| \, 1)\} \tag{A7}$$

$$\leq \sum_{j=1}^{p} D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_j} \| \, 1) \tag{A8}$$

which is the desired upper-bound.

Appendix B

PROOF OF THE LINK BETWEEN THE OPTIMIZATION OF THE DIVERGENCE AND THE CSP SOLUTION

The fact that any Rayleigh quotient is bounded by the maximum and minimum eigenvalues of the associated matrix pencil

$$\lambda_1 \leq \frac{\boldsymbol{w}_i^T \mathbf{P} \boldsymbol{w}_i}{\boldsymbol{w}_i^T \mathbf{Q} \boldsymbol{w}_i} \leq \lambda_n \tag{A1}$$

can be used to recursively prove that the minimax value of the divergence is equal to

$$\begin{aligned} &\min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T \mathbf{P} \boldsymbol{w}_i \parallel \kappa\, \boldsymbol{w}_i^T \mathbf{Q} \boldsymbol{w}_i) \\ &\quad = \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} D_{AB}^{(\alpha,\beta)}(\frac{\boldsymbol{w}_i^T \mathbf{P} \boldsymbol{w}_i}{\boldsymbol{w}_i^T \mathbf{Q} \boldsymbol{w}_i} \parallel \kappa) \qquad \text{(A2)} \\ &\quad = D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_i'} \parallel \kappa), \qquad \text{(A3)} \end{aligned}$$

where permutation π' sorts the divergence of the eigenvalues from κ in descending order

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_1'} \parallel \kappa) \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_2'} \parallel \kappa) \geq \cdots \geq D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_n'} \parallel \kappa). \tag{A4}$$

The minimax value is then attained for the eigenvectors

$$\boldsymbol{v}_{\pi_i'} = \arg \min_{\dim\{\mathcal{W}\}=n-i+1} \max_{\boldsymbol{w}\in\mathcal{W}} D_{AB}^{(\alpha,\beta)}(\boldsymbol{w}_i^T \mathbf{P} \boldsymbol{w}_i \parallel \kappa\, \boldsymbol{w}_i^T \mathbf{Q} \boldsymbol{w}_i), \qquad i = 1, \ldots, p. \tag{A5}$$

For the coincidence of the set of solutions $\{\boldsymbol{v}_{\pi_1'}, \ldots, \boldsymbol{v}_{\pi_p'}\}$ in (A5) with the set of spatial filters $\{\boldsymbol{v}_1^{(c_1)}, \ldots, \boldsymbol{v}_k^{(c_1)}, \boldsymbol{v}_{n-(p-k)+1}^{(c_1)}, \ldots, \boldsymbol{v}_n^{(c_1)}\}$ that define the $\mathbf{W}_{CSP}$, the eigenvalues $\lambda_{\pi_1}, \ldots, \lambda_{\pi_p}$ that maximize their divergence from κ, should all belong to the upper and lower sets of eigenvalues defined in (6.51). For this to be true, it necessary and sufficient that the divergence of the last selected eigenvalue λ_{π_p} from κ upper-bounds with inequality all the divergences between an inner eigenvalue λ_i and κ, in the sense that

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_p} \parallel \kappa) > \max_{i\in[k+1\ ,\ n-(p-k)]} D_{AB}^{(\alpha,\beta)}(\lambda_i \parallel \kappa) \tag{A6}$$

The domain of κ for which this strict inequality holds true is

$$\kappa \in (\kappa_{\inf}\ ,\ \kappa_{\sup}) \tag{A7}$$

where the bounds

$$\kappa_{\inf} \equiv \mathcal{K}(\lambda_{k+1}\ ,\ \lambda_{n-(p-k)+1\rfloor}) \tag{A8}$$

$$\kappa_{\sup} \equiv \mathcal{K}(\lambda_{k}\ ,\ \lambda_{n-(p-k)}) \tag{A9}$$

respectively equalize the value of the divergences

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_p}\|\ \kappa_{\inf}) = D_{AB}^{(\alpha,\beta)}(\lambda_{k+1}\|\ \kappa_{\inf}) = D_{AB}^{(\alpha,\beta)}(\lambda_{n-(p-k)+1}\|\ \kappa_{\inf}) \tag{A10}$$

and

$$D_{AB}^{(\alpha,\beta)}(\lambda_{\pi_p}\|\ \kappa_{\sup}) = D_{AB}^{(\alpha,\beta)}(\lambda_{k}\|\ \kappa_{\sup}) = D_{AB}^{(\alpha,\beta)}(\lambda_{n-(p-k)}\|\ \kappa_{\sup}). \tag{A11}$$

Appendix C

DIFFERENTIAL OF THE INVERSE SQUARE ROOT OF A SPD MATRIX

Let $\mathbf{X}$ be any symmetric possitive definite matrix (SPD). We would like to obtain the differential of its inverse square root $d\mathbf{X}^{-\frac{1}{2}}$ in terms of the matrix $\mathbf{X}$ and its differential $d\mathbf{X}$, and later use this result to simplify the desired expression $d\mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}$. For this purpose, we start from the trivial identity $\mathbf{I}_p = \mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}$ and take differentials on both sides of this equality, with the help of the product rule for differentials we obtain

$$\mathbf{0} = d\mathbf{I}_p = d(\mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}) = d\mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} + \mathbf{X}^{-\frac{1}{2}}d\mathbf{X}^{\frac{1}{2}}. \tag{A1}$$

Solving for the differential

$$d\mathbf{X}^{-\frac{1}{2}} = -\mathbf{X}^{-\frac{1}{2}}d\mathbf{X}^{\frac{1}{2}}\mathbf{X}^{-\frac{1}{2}}, \tag{A2}$$

we see it as a function of $\mathbf{X}$ and $d\mathbf{X}^{\frac{1}{2}}$. Then, we simplify $d\mathbf{X}^{\frac{1}{2}}$ with the help of the another trivial identify $\mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}})^T = \mathbf{X}$. We take again differentials on both sides of the equality

$$d(\mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}})^T) = d\mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}})^T + \mathbf{X}^{\frac{1}{2}}(d\mathbf{X}^{\frac{1}{2}})^T = d\mathbf{X} \tag{A3}$$

and obtain the special solution

$$d\mathbf{X}^{\frac{1}{2}} = \tfrac{1}{2}d\mathbf{X}(\mathbf{X}^{-\frac{1}{2}})^T. \tag{A4}$$

The substitution of (A4) in (A3) yields the differential of the inverse symmetric square root of the SPD matrix

$$d\mathbf{X}^{-\frac{1}{2}} = -\mathbf{X}^{-\frac{1}{2}}\left(\tfrac{1}{2}d\mathbf{X}(\mathbf{X}^{-\frac{1}{2}})^T\right)\mathbf{X}^{-\frac{1}{2}}. \tag{A5}$$

Finally, by the symmetry of $\mathbf{X}^{-\frac{1}{2}}$, we prove the desired result

$$d\mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} = -\frac{1}{2}\mathbf{X}^{-\frac{1}{2}}d\mathbf{X}\,(\mathbf{X}^{-\frac{1}{2}})^T = -\frac{1}{2}\mathbf{X}^{-\frac{1}{2}}d\mathbf{X}\,\mathbf{X}^{-\frac{1}{2}} \tag{A6}$$

Appendix D

THE GRADIENT OF THE KL DIVERGENCE BETWEEN GAUSSIAN DENSITIES

The Kullback–Leibler (KL) divergence between the Gaussian densities $p(\boldsymbol{x}|c_2)$ and $p(\boldsymbol{x}|c_1)$, of zero mean and with respective covariance matrices $Cov(\mathbf{Y}|c_1) = \mathbf{W}^T\mathbf{P}\mathbf{W}$ and $Cov(\mathbf{Y}|c_2) = \mathbf{W}^T\mathbf{Q}\mathbf{W}$, is equal to

$$\begin{aligned} Div_{KL}(p(\boldsymbol{x}|c_2) \,\|\, p(\boldsymbol{x}|c_1)) &= \tfrac{1}{2}\log|\mathbf{W}^T\mathbf{P}\mathbf{W}| - \tfrac{1}{2}\log|\mathbf{W}^T\mathbf{Q}\mathbf{W}| \\ &\quad + \tfrac{1}{2}\,\mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W}) - \mathbf{I}_p\}. \end{aligned} \tag{A1}$$

This subsection explains the operations involved in obtaining its gradient. The first differential of the log-determinant terms is

$$d\log|\mathbf{W}^T\mathbf{P}\mathbf{W}| = \mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} d(\mathbf{W}^T\mathbf{P}\mathbf{W})\} \tag{A2}$$

$$= \mathrm{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(d\mathbf{W}^T\,\mathbf{P}\mathbf{W} + \mathbf{W}^T\,\mathbf{P}\,d\mathbf{W})\} \tag{A3}$$

$$= 2\,\mathrm{tr}\{[\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}]\, d\mathbf{W}^T\}. \tag{A4}$$

By using the relationship between the first differential and the gradient

$$d\log|\mathbf{W}^T\mathbf{P}\mathbf{W}| = \mathrm{tr}\{[\nabla_{\mathbf{W}}\log|\mathbf{W}^T\mathbf{P}\mathbf{W}|]\, d\mathbf{W}^T\} \tag{A5}$$

one can identify from (A4) that

$$\nabla_{\mathbf{W}}\,\tfrac{1}{2}\log|\mathbf{W}^T\mathbf{P}\mathbf{W}| = \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \tag{A6}$$

and, similarly,

$$\nabla_{\mathbf{W}}\,[-\tfrac{1}{2}\log|\mathbf{W}^T\mathbf{Q}\mathbf{W}|] = -\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}\,. \tag{A7}$$

On the other hand, the first differential of the trace term simplifies to

$$\begin{aligned} d\left[\tfrac{1}{2}\operatorname{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})-\mathbf{I}_p\}\right] &= \tfrac{1}{2}\operatorname{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}d(\mathbf{W}^T\mathbf{Q}\mathbf{W})\} \\ &+ \tfrac{1}{2}\operatorname{tr}\{d(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})\} \\ &= \tfrac{1}{2}\operatorname{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}d(\mathbf{W}^T\mathbf{Q}\mathbf{W})\} \\ &- \tfrac{1}{2}\operatorname{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}d(\mathbf{W}^T\mathbf{P}\mathbf{W})(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \\ &(\mathbf{W}^T\mathbf{Q}\mathbf{W})\} \\ &= \tfrac{1}{2}\operatorname{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(d\mathbf{W}^T\mathbf{Q}\mathbf{W}+\mathbf{W}^T\mathbf{Q}d\mathbf{W})\} \\ &- \tfrac{1}{2}\operatorname{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(d\mathbf{W}^T\mathbf{P}\mathbf{W}+\mathbf{W}^T\mathbf{P}d\mathbf{W}) \\ &\times(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})\} \\ &= \tfrac{1}{2}\operatorname{tr}\{2\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}\,d\mathbf{W}^T\} \\ &- \tfrac{1}{2}\operatorname{tr}\{2\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \\ &d\mathbf{W}^T\}. \end{aligned} \tag{A8}$$

From which one can also identify

$$\begin{aligned} \nabla_{\mathbf{W}}\ \left[\tfrac{1}{2}\operatorname{tr}\{(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})-\mathbf{I}_p\}\right] &= \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \\ &- \mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}. \end{aligned} \tag{A9}$$

Once we have obtained in (A6), (A7) and (A9) the gradients of the partial terms that are involved in the definition (A1) of the KL divergence, their simple addition yields the complete gradient of the KL divergence with respect to $\mathbf{W}$, which is given by

$$\begin{aligned} \tfrac{\partial}{\partial\mathbf{W}}Div_{KL}(p(\boldsymbol{x}|c_2)\,\|\,p(\boldsymbol{x}|c_1)) &= -\mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1}+\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1} \\ &+ \mathbf{Q}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}-\mathbf{P}\mathbf{W}(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}(\mathbf{W}^T\mathbf{Q}\mathbf{W})^{-1} \\ &(\mathbf{W}^T\mathbf{P}\mathbf{W})^{-1}. \end{aligned} \tag{A10}$$